Посвящается Рут

Джудиа Перл, Дана Маккензи

ДУМАЙ «ПОЧЕМУ?»

ПРИЧИНА И СЛЕДСТВИЕ КАК КЛЮЧ К МЫШЛЕНИЮ

Издательство АСТ
Москва

УДК 159.95
ББК 88.3
П26

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

Печатается с разрешения правообладателя Brockman, Inc.

Перл, Джудиа.

П26 Думай «почему?». Причина и следствие как ключ к мышлению / Джудиа Перл, Дана Маккензи; перевод с английского Т. Мамедовой, М. Антипова. — Москва: Издательство АСТ, 2022. — 448 с.

ISBN 978-5-17-123140-8 (Trend book)
ISBN 978-5-17-146449-3 (Власть и успех)

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.

УДК 159.95
ББК 88.3

ISBN 978-5-17-123140-8
(Trend book)
ISBN 978-5-17-146449-3
(Власть и успех)

Содержание

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 путешествиях, 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 степени. 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 предотвращения 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 *XX*I 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 науч-

ные 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 препятствием, 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 рассказывали 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 *X*VII века люди вполне удовлетворялись своей способностью справляться с неопределенностью в повседневной жизни — от перехода 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 сожа-

лению, 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 названием «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 принимать, 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 причинности 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 эконометри-

ке) 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 получили 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 вероятность 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 надежные 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 безошибочными 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 продолжить 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», которое 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 оцениваемой 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 со-

стоянии предсказать, что пациент со свойством *Z* проживет L лет. Но теперь ее перевели в новую больницу в другой 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 инструментом. 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 по-прежнему во многом определяют математическое основания для диаграмм причинности. Помимо постепенного знакомства 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 отвлечься 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 душе. 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 были 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. Люди 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 практически 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.

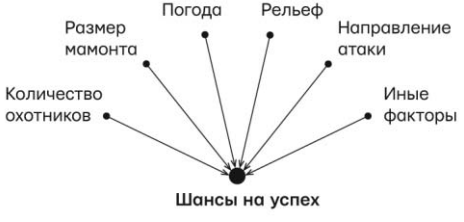


Рис. 2. Предполагаемые причины успеха в охоте на мамонта

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 менталь-

ной 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.

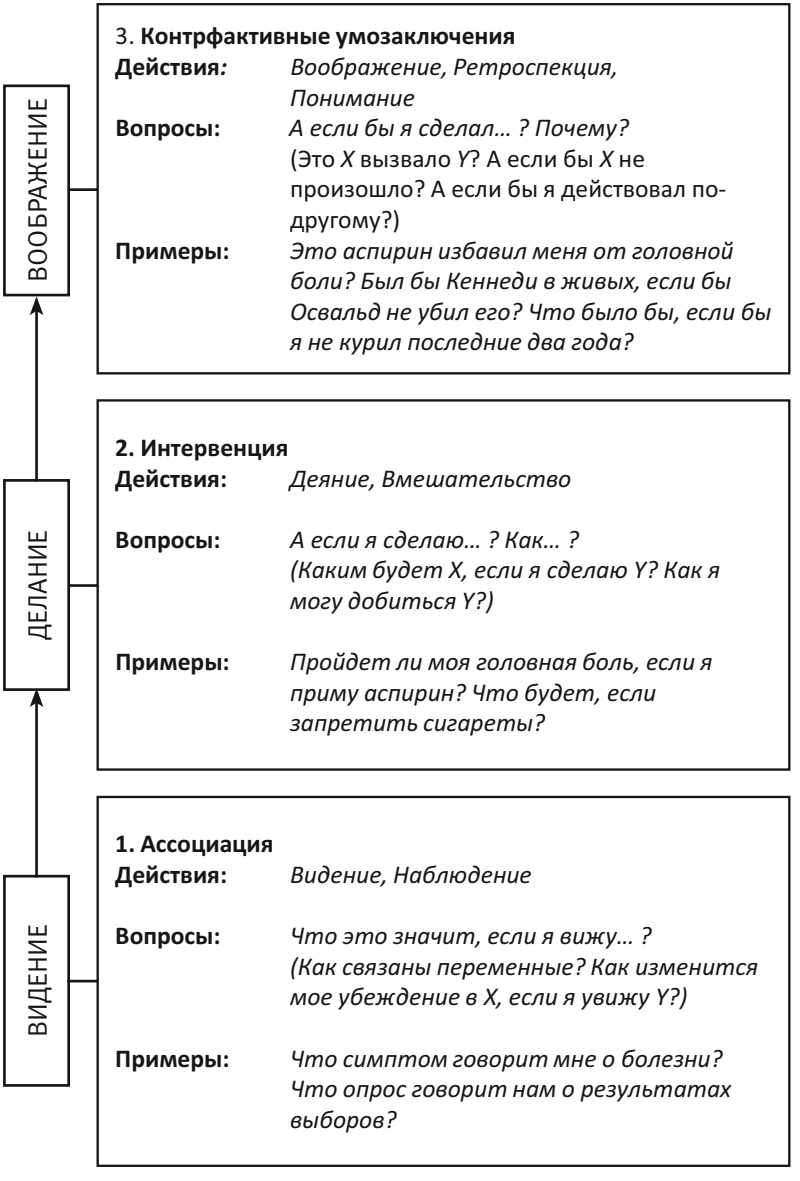


Рис. 3. Лестница Причинности с представляющими ее организмами на каждом уровне. Большинство животных, так же как и сегодняшние обучающиеся машины, находятся на первой перекла-

дине — 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 глубинного 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 версий 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 всеми 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 действия 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 разработать 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 быстро 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. Важнейшим 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*, что переменная *X* равна 1, и *сделаем* переменную *Y* равной 0, а переменную *Z* равной 1?». Если бы переменных было больше или если бы у каждой было свыше двух состояний, то число возможностей вышло бы за пределы нашего воображения. В список Сёрла пришлось бы внести пунктов больше, чем атомов во Вселенной. Очевидно, что простой список вопросов и ответов никогда не сымитирует интеллект ребенка, не говоря уже об интеллекте взрослого.

Человеческому мозгу необходимы компактное представление информации, а также эффективная процедура, которая позволит должным образом интерпретировать каждый вопрос и вычленить нужный ответ из этого сохраненного представления. Таким образом, чтобы пройти мини-тест Тьюринга, нужно

снабдить машины такой же эффективной репрезентацией и алгоритмом для получения ответа.

Эта репрезентация не просто существует, она по-детски проста — я говорю о диаграмме причинности. Мы уже видели один пример — диаграмму об охоте на мамонта. С учетом невероятной легкости, с какой люди могут передавать свои знания в диаграммах из стрелок и точек, я верю, что у нас в мозге действительно существует такая репрезентация. Но, что важнее для наших целей, эти модели позволяют пройти мини-тест Тьюринга, тогда как ни одна другая модель на это не способна. Давайте рассмотрим некоторые примеры.

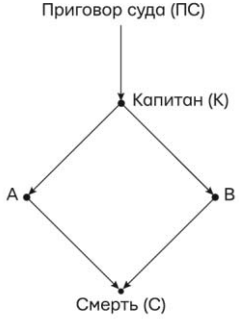


Рис. 4. Диаграмма причинности для примера с расстрелом. A и *B* представляют действия солдат A и *B*

Предположим, что расстрельная команда собирается казнить узника. Чтобы это произошло, должна случиться определенная последовательность событий. Сначала суд выносит приговор о расстреле. Его доводят до капитана, который дает сигнал солдатам из расстрельной команды (*A* и *B*) стрелять. Будем считать, что они послушные исполнители и опытные снайперы, поэтому действуют только по команде, и если один из них выстрелит, то узник умрет.

На рис. 4 показана диаграмма, представляющая сюжет, который я только что изложил. Каждое из неизвестных (*ПС*, *К*, *A*, *B*, *C*) является переменной со значением «верно/неверно».

Например, «*С* = верно» свидетельствует, что узник мертв; «*С* = неверно» выражает, что узник жив. «*ПС* = неверно» означает, что приговор не был вынесен; «*ПС* = верно» — что он был вынесен и т.д.

Диаграмма позволяет нам отвечать на вопросы о причинах, соответствующие разным уровням Лестницы. Во-первых, можно ответить на вопросы о связях (т.е. о том, что один факт говорит нам о другом). Если узник мертв, значит ли это, что приговор был вынесен? Мы (или компьютер) способны изучить диаграмму, проследить правила, стоящие за каждой стрелкой и, используя стандартную логику, прийти к выводу, что два солдата не выстрелили бы без команды капитана. Подобным образом капитан не дал бы команды, если бы в его распоряжении не было приговора. Поэтому ответ на наш вопрос — да. Другой вариант: предположим, мы узнали, что выстрелил *А*. Что это говорит нам о действиях *В*? Следуя стрелкам, компьютер приходит к выводу, что В тоже должен был выстрелить (А не стал бы стрелять, если бы капитан не дал сигнала, значит, *В* точно стрелял). Это справедливо, даже когда *А* не вызывает *В* (между *А* и *В* нет стрелки).

Поднимаясь по Лестнице Причинности, можно поставить вопрос об интервенции. А если солдат *А* по собственной инициативе решит выстрелить, не дожидаясь команды капитана? Будет ли узник жив или мертв? Вообще, этот вопрос сам по себе содержит некоторое противоречие. Я сейчас сказал вам, что *А* выстрелит, только если получит команду, а теперь мы спрашиваем, что будет, если он выстрелит без команды. Если просто использовать правила логики, как обычно делают компьютеры, этот вопрос становится бессмысленным. Как говорил в таких случаях робот из телесериала 1960-х годов «Затерянные в космосе», «это не вычисляется».

Если мы хотим, чтобы наш компьютер понимал причинно-следственные связи, нужно научить его нарушать правила. Он должен усвоить, что просто наблюдать за событием и быть его причиной — разные вещи. Мы говорим компьютеру: «Во всех случаях, когда ты становишься причиной события, убери все стрелки, указывающие на это событие, и продолжай ана-

лиз с помощью обычной логики, как будто стрелок никогда не было». Таким образом, мы стираем все стрелки, ведущие к переменной, ставшей объектом интервенции (*A*). Также мы вручную настраиваем эту переменную, присваивая ей значение («верно»). Обоснование для этой странной «хирургической операции» простое: вызывая событие к жизни, мы освобождаем его от всех других влияющих обстоятельств и подвергаем только одному — тому, которое заставляет его случиться.

На рис. 5 показана диаграмма причинности на основе нашего примера. Эта интервенция неизбежно приводит к смерти узника. Такова причинная функция стрелки, ведущей от *A* к *C*.

Рис. 5. Рассуждение об интервенциях. Солдат А решает выстрелить; стрелка от К к А стерта, и А получает значение «верно».

Заметим, что этот вывод согласуется с нашим интуитивным суждением: выстрел *A*, сделанный без команды, приведет к смерти узника, потому что хирургическое вмешательство оставило стрелку от *A* к *C* неприкосновенной. Кроме того, мы придем к выводу, что *B* (по всей вероятности) не выстрелил; ничего, связанное с решением *A*, не должно влиять на переменные в модели, не являющиеся результатом выстрела *A*. Это утверждение стоит повторить. Если мы *видим*, что А стреляет, то делаем вывод, что *B* тоже выстрелил. Но если *A решает* выстрелить или если мы *заставляем A* выстрелить, то верно обратное. В этом разница между тем, чтобы *видеть*, и тем,

чтобы *делать*. Только компьютер, способный уловить эту разницу, может пройти мини-тест Тьюринга.

Заметим, что, если бы мы просто собирали большие данные, это не помогло бы подняться по Лестнице и ответить на вопросы, заданные выше. Предположим, вы журналист, который ежедневно собирает информацию о расстрелах. В ваших данных будут только два типа событий: либо все пять переменных верны, либо все они неверны. Вот почему, располагая данными такого рода и не понимая, кто кого «слушает», вы (или любой алгоритм машинного обучения) ни за что не предскажете, что будет, если убедить снайпера A не стрелять.

Наконец, чтобы проиллюстрировать третий уровень Лестницы Причинности, давайте зададим контрфактивный вопрос. Предположим, мертвый узник лежит на земле. Из этого мы можем сделать вывод (используя первый уровень), что *A* выстрелил, *B* выстрелил, капитан подал сигнал, а суд вынес приговор. А если бы *A* решил не стрелять? Остался бы узник в живых? Этот вопрос требует от нас сравнения реального мира с вымышленным и противоречащим нашему, в котором *A* не выстрелил. В этом вымышленном мире стрелка, ведущая к *A*, стерта, чтобы *A* мог не слушать *K*. Переменной *A* присвоено значение «неверно», но ее предыдущая история остается той же, что и в реальном мире. Итак, вымышленный мир выглядит как на рис. 6.

Чтобы пройти этот мини-тест Тьюринга, наш компьютер должен прийти к выводу: узник будет мертв в вымышленном мире тоже, потому что там его убил бы выстрел *B*, т.е. доблестный отказ *A* не спас бы его жизни. Несомненно, по этой единственной причине и существуют расстрельные команды: они гарантируют, что приговор будет приведен в исполнение, и снимают некоторое бремя ответственности с каждого стрелка в отдельности: все они могут с чистой (относительно) совестью утверждать, что их действия не привели к смерти узника, потому что «он все равно бы умер».

Может показаться, что мы приложили массу усилий, стараясь ответить на ненастоящие вопросы, с которыми и так все было ясно. Я полностью согласен! Рассуждения о причинно-след-

ственных связях даются вам без труда, потому что вы человек, и когда-то вам было три года, и у вас был замечательный трехлетний мозг, который понимал причинно-следственные связи лучше, чем любое животное или компьютер. Весь смысл мини-теста Тьюринга в том, чтобы рассуждения о причинности стали по силам и машинам. В ходе этого процесса мы могли узнать что-то новое о том, как это делают люди. Все три примера показывают, что компьютеры нужно научить выборочно нарушать правила логики. Компьютерам трудно это делать, а детям очень легко. (И пещерным людям тоже! Человекольва не создали бы, не нарушив правила о том, какая голова подходит для того или иного тела.)

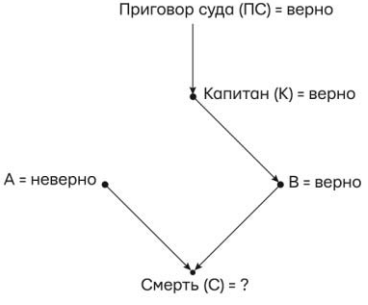


Рис. 6. Контрфактивное рассуждение. Мы наблюдаем, что узник мертв и спрашиваем, что случилось бы, если бы солдат А решил не стрелять.

Но все же не будем почивать на лаврах, утверждаясь в человеческом превосходстве. В очень многих ситуациях людям, скорее всего, будет гораздо сложнее прийти к верным выводам о причинно-следственных связях. Так, может возникнуть гораздо больше переменных и они окажутся не просто бинарными (верно/неверно). Вместо того чтобы гадать, жив или мертв узник, нам, предположим, понадобится предсказать, насколько вырастит безработица, если поднять минимальную заработную плату. Такого рода количественное рассуждение о причинно-следственных связях обычно не под силу нашей

интуиции. Кроме того, в примере с расстрельной командой мы исключили неопределенность: скажем, капитан дал команду через долю секунды после того, как солдат *A* решил выстрелить или у солдата *B* заклинило ружье и т.д. Чтобы справиться с неопределенностью, нам нужна информация о вероятности таких ненормальных ситуаций.

Позвольте привести пример, в котором от вероятностей зависит все. Он отражает споры, разгоревшиеся в Европе, когда впервые появилась вакцина от оспы. Тогда статистические данные неожиданно показали, что от прививки умирает больше людей, чем от самой болезни. Естественно, некоторые люди использовали эту информацию как аргумент в пользу запрета прививок, тогда как на деле она спасала жизни, избавляя от риска заболеть. Давайте рассмотрим вымышленные данные, чтобы проиллюстрировать этот эффект и разрешить спор.

Представим, что из миллиона детей 99% получает прививку, а 1% — нет. Если ребенок привит, то у него или у нее есть один шанс из 100 на побочную реакцию, и в одном случае из 100 реакция может стать смертельной. В то же время, если ребенок не прививается, у него или у нее очевидно нет риска получить побочную реакцию на прививку, однако есть один шанс из 50 заболеть оспой. Наконец, давайте считать, что оспа смертельна в одном случаев из пяти.

Я думаю, вы согласитесь, что вакцинация — хорошая мысль. Шансы получить побочную реакцию ниже, чем шансы заразиться оспой, и сама реакция гораздо менее опасна, чем болезнь. Но давайте посмотрим на данные. Из миллиона детей 990 тысяч получают прививку, у 9 900 возникает побочная реакция и 99 умирает. В то же время 10 тысяч не прививаются, 200 заражаются оспой и 40 умирает. В результате от вакцины умирает больше детей (99), чем от болезни (40).

Я понимаю родителей, которые готовы устроить демонстрацию перед министерством здравоохранения с лозунгами «Прививки убивают!». И вроде бы данные подтверждают их позицию — прививки действительно вызывают больше смертей, чем сама оспа. Но на их ли стороне логика? Надо ли запретить прививки или же стоит взять в расчет предотвра-

щенные смерти? На рис. 7 вы найдете диаграмму причинности для этого примера.

Когда мы начали, вакцинировалось 99% детей. Теперь мы задаем контрфактивный вопрос: «А что, если снизить число вакцинированных до нуля?». Используя вероятности, которые я привел выше, мы можем прийти к выводу, что из миллиона детей 20 тысяч заразились бы оспой и 4 тысячи умерли бы. Сравнивая контрфактивный мир с настоящим, мы видим, что отсутствие прививок стоило бы жизни 3 861 ребенку (разница между 4 тысячами и 139). Стоит поблагодарить язык контрфактивных суждений, который помогает нам избежать таких потерь.

Главный урок для изучающих причинность состоит в том, что модель причинности подразумевает гораздо больше, чем простое рисование стрелок. За стрелками стоят вероятности. Когда мы рисуем стрелку от *X* к *Y*, мы подразумеваем, что некоторое правило или функция, определяющие вероятность, указывают, как изменится *Y*, если изменится *X*. В некоторых случаях мы знаем правило, но вероятнее, что его придется вывести из данных. Одна из самых интригующих особенностей Революции Причинности, однако, состоит в том, что во многих случаях можно оставить математические данные абсолютно неопределенными. Очень часто *структура самой диаграммы* позволяет нам оценить самые разные причинные и контрфактивные отношения — простые или сложные, детерминистские или вероятностные, линейные или нелинейные.

С вычислительной точки зрения наша схема для мини-теста Тьюринга также примечательна тем, что мы использовали один порядок действий для всех трех примеров: перевели историю в диаграмму, выслушали запрос, сделали «хирургическое вмешательство», соответствующее конкретному запросу (интервенционное или контрфактивное; если запрос о связях, вмешательства не требуется), использовали измененную причинную модель, чтобы вычислить ответ. Нам не пришлось обучать машину множеству новых запросов каждый раз, когда история менялась. Этот подход достаточно гибкий, чтобы работать каждый раз, когда возможно нарисовать диаграмму

причинности — применительно к мамонтам, расстрельным командам или прививкам. Именно это мы и хотим получить от механизма причинного вывода — именно такой гибкостью обладаем мы, люди.

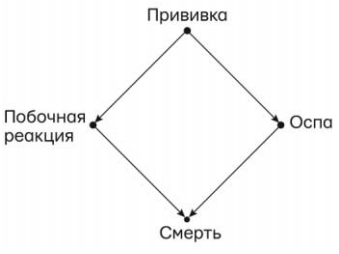


Рис. 7. Диаграмма причинности для примера с прививками. Полезна ли вакцинация?

Конечно, в самой диаграмме нет ничего волшебного. Она позволяет достичь успеха, потому что содержит информацию о причинах; т.е., составив диаграмму, мы спросили: «Кто может вызвать смерть заключенного напрямую?» или «Каков непосредственный эффект от вакцинации?». Если бы мы составляли диаграмму, спрашивая исключительно об ассоциациях, она не дала бы нам таких возможностей. Например, если бы на рис. 7 мы направили стрелку от оспы к прививкам, то получили бы такие же связи между данными, но пришли бы к ошибочному мнению о том, что оспа влияет на вакцинацию.

Но давайте внимательнее рассмотрим этот критерий повышения вероятности и увидим, где он дает сбой. Вопрос общей причины или *вмешивающегося фактора* для *X* и *Y* доставлял философам максимум неприятностей. Если взять критерий повышения вероятности как таковой, то придется заключить, что продажи мороженого вызывают преступления, так как вероятность преступлений выше в месяцы, когда продается больше мороженого. В этом конкретном случае мы объясним феномен тем, что и продажи мороженого, и преступность выше летом, когда погода теплее. Тем не менее у нас все равно

остается вопрос: какой общий философский критерий способен определить, что причина — погода, а не продажи мороженого?

Философы изо всех сил старались исправить это определение — они учли в нем так называемые фоновые факторы (еще одно название для осложняющих факторов) и привлекли критерий $P(Y \mid X, K = k) > P(Y \mid K = k)$, где K обозначает некие фоновые переменные. Более того, этот критерий работает для нашего примера с мороженым, если считать температуру фоновой переменной. Скажем, если мы рассмотрим только дни, когда температура достигает 30 °C ($K = 30$), то не найдем остаточных связей между мороженым и преступлениями. Иллюзия, что вероятность повышается, возникнет, только если мы сравним дни, когда было +30 °C, с днями, когда был 0 °C.

И все же ни один философ не смог дать убедительный общий ответ на вопрос: какие переменные необходимо включить в набор общих переменных K и сделать условием задачи? Проблема очевидна: осложняющие переменные — это тоже понятие из сферы причинности, поэтому они не поддаются описанию с точки зрения вероятности. В 1983 году Нэнси Картрайт вышла из тупика и обогатила описание фонового контекста элементами причинности. Она предложила учитывать только факторы, «причинно релевантные» для следствия. Позаимствовав это понятие со второго уровня Лестницы Причинности, она, по сути дела, отказалась от идеи определять причины на основе исключительно вероятности. Это был прогресс, но критики получили возможность утверждать, что мы определяем причину через нее саму.

Философские споры по поводу подобающего содержания K продолжались более 20 лет и зашли в тупик. Замечу, что мы увидим верный критерий в главе 4 и я не буду портить здесь сюрприз. На данный момент достаточно сказать, что это критерий практически нереально сформулировать без диаграмм причинности.

Обобщая, следует сказать, что вероятностная причинность всегда сталкивалась с осложняющими переменными. Каждый раз, когда приверженцы вероятностной причинности пытаются починить корабль, снабдив его новым корпусом, он натыкается

на тот же подводный камень и получает очередную протечку. Но, если выразить «рост вероятности» на языке условных вероятностей, как ни подлатывай корпус, на следующий уровень Лестницы не попадешь. Как бы странно это ни звучало, понятие повышения вероятности нельзя объяснить в терминах вероятностей.

Верный способ спасти идею повышения вероятности — использовать оператор *do*: можно сказать, что *X* вызывает *Y*, если $P(Y \mid do(X)) > P(Y)$. Поскольку интервенция — понятие второго уровня, это определение способно отразить причинную интерпретацию повышения вероятности, а еще оно будет работать на диаграммах причинности. Другими словами, если у нас на руках диаграмма причинности и данные, и исследователь спрашивает, действительно ли $P(Y \mid do(X)) > P(Y)$, мы в состоянии дать связный алгоритмический ответ и таким образом решить, является ли *X* причиной *Y* в плане повышения вероятности.

Обычно я обращаю много внимания на то, что философы хотят сказать о скользких понятиях, таких как причинность, индукция или логика научных рассуждений. У философов есть преимущество: они стоят в стороне от оживленных научных дебатов и от реалий взаимодействия с данными на практике. Они в меньшей степени, чем другие ученые, заражены антипричинными предубеждениями статистики.

Они могут привлечь традицию восприятия причинности, которая восходит к Аристотелю, и говорить о причинности, не краснея и не пряча ее за этикеткой «ассоциации».

Однако, стараясь перевести понятие причинности на язык математики, что само по себе идея, достойная похвалы, философы слишком быстро прибегли к единственному известному им языку, который может описать неопределенность, — к языку вероятности. За последний десяток лет они в основном преодолели это заблуждение, но, к несчастью, похожие идеи сейчас рассматриваются в эконометрике под названиями вроде «причинность по Грэнджеру» и «векторная автокорреляция».

И сейчас я сделаю признание: я совершил ту же ошибку. Я не всегда ставил причинность на первое место, а вероят-

ность — на второе. Наоборот! Когда я стал работать над искусственным интеллектом в начале 1980-х годов, я думал, что неопределенность — самая важная вещь, которой не хватает ИИ. Более того, я настаивал на том, чтобы неопределенность была представлена с помощью вероятностей. Таким образом, как я объясняю в главе 3, я разработал подход к рассуждениям в условиях неопределенности под названием «байесовские сети», который имитирует, как идеализированный, децентрализованный мозг может включить вероятности в принятие решений. Если мы видим определенные факты, байесовские сети способны быстро вычислить вероятность верности или неверности определенных фактов. Неудивительно, что байесовские сети сразу обрели популярность в сообществе ИИ и даже сегодня считаются ведущей парадигмой в искусственном интеллекте для рассуждений при неопределенности.

Хотя продолжающийся успех байесовских сетей чрезвычайно радует меня, они не смогли закрыть зазор между искусственным и человеческим интеллектом. Я уверен, что вам понятно, какой составляющей не хватает — причинности. Да, призраки причинности в изобилии витали рядом. Стрелки неизменно вели от причин к следствиям, и практики часто замечали, что диагностические системы становятся неуправляемыми, если направление стрелок меняется в обратную сторону. Но по большей части мы думали, что эта культурная привычка — артефакт былых сценариев мышления, а не центральный аспект разумного поведения.

В то время меня так опьянила сила вероятностей, что я счел причинность второстепенным понятием — просто удобством или ментальной скорописью для выражения вероятностных зависимостей и отделения релевантных переменных от нерелевантных.

В своей книге 1988 года «Вероятностные рассуждения в интеллектуальных системах» (*Probabilistic Reasoning in Intelligent Systems*) я писал: «Причинность — язык, на котором мы можем эффективно обсуждать определенные структуры в отношениях релевантности». Я смущаюсь, вспоминая эти слова сегодня, потому что релевантность — очевидно, понятие первого уров-

ня. Еще ко времени, когда книга была напечатана, в глубине души я знал, что был неправ. Для моих коллег — специалистов по компьютерным наукам книга стала библией вероятностных рассуждений в условиях неопределенности, но я уже чувствовал себя еретиком.

Байесовские сети существуют в мире, где все вопросы сводятся к вероятностям или (в терминах этой главы) степеням связи между переменными; они не могли подняться на второй или третий уровни Лестницы Причинности. К счастью, потребовалось всего два небольших изменения, чтобы забраться наверх. Сначала, в 1991 году, благодаря идее сделать графику «хирургическую операцию», получилось применить его и к наблюдениям, и к интервенциям. Еще один поворот, в 1994 году, вывел их на третий уровень — они стали применимы к контрфактивным суждениям. Но все это заслуживает обсуждения ниже. Главное в следующем: в то время как вероятности кодируют наши представления о статичном мире, причинность говорит нам, как вероятности меняются (и меняются ли) в статичном мире, будь то посредством интервенции или воображения.

Глава 2
От государственных пиратов до морских свинок: становление причинного вывода

И всё-таки она вертится.
Приписывается Галилео Галилею,
1564—1642

Почти два столетия одним из самых постоянных ритуалов в британской науке были вечерние лекции по пятницам в Королевском институте Великобритании в Лондоне. Многие великие открытия *XIX* столетия впервые были представлены публике именно там: принципы фотографии Майкла Фарадея в 1839-м; электроны в докладе Джозефа Джона Томсона в 1897-м; сжижение водорода в лекции Джеймса Дьюара в 1898-м.

Зрелищности на этих мероприятиях всегда придавали большое значение: здесь наука буквально становилась театром, и зрители, сливки британского общества, были разодеты в пух и прах (мужчины непременно в смокингах с черными галстуками). С боем часов вечернего докладчика почтительно вводили в аудиторию. По традиции он начинал лекцию тотчас же, без представления или вступления. Эксперименты и наглядные демонстрации были частью зрелища.

Вечером 9 февраля 1877 года докладчиком был Фрэнсис Гальтон, член Королевского общества, двоюродный брат Чарлза Дарвина, известный исследователь Африки, изобретатель дактилоскопии и классический пример ученого джентльмена

викторианской эпохи. Название доклада Гальтона гласило: «Типичные законы наследственности». Экспериментальный прибор, сделанный им для доклада, представлял собой странное устройство, которое он назвал квинкунксом (сейчас его часто именуют доской Гальтона). Похожее приспособление используется в американской телевикторине «Цена верна». Доска Гальтона состояла из рядов воткнутых в дерево булавок, расположенных таким образом, что любые три соседние булавки образовывали равносторонний треугольник; через отверстие сверху можно было насыпать маленькие металлические шарики, которые, ударяясь о булавки, как в пинболе, скатывались вниз, в итоге попадая в один из пазов внизу доски (см. фронтиспис). Для каждого индивидуального шарика отскоки влево и вправо от булавок по мере скатывания вниз распределяются совершенно случайно. Однако если в устройство Гальтона всыпать много шариков, становится видна удивительная закономерность: накопившиеся на дне шарики всегда образуют грубое подобие колоколообразной кривой. Пазы ближе к центру будут содержать больше всего шариков, а по мере продвижения к обоим краям доски число шариков в каждом пазу будет постепенно уменьшаться.

У такого распределения есть математическое объяснение. Путь каждого отдельного шарика подобен последовательности независимых подбрасываний монеты. Всякий раз, когда шарик сталкивается с булавкой, он отскакивает или вправо, или влево, и со стороны его движение кажется совершенно случайным. Сумма результатов — число отскакиваний вправо относительно числа отскакиваний влево — определяет, в каком из пазов шарик закончит свой путь. Согласно центральной предельной теореме теории вероятностей, доказанной в 1810 году Пьером Симоном Лапласом, любой подобный случайный процесс, эквивалентный большому числу последовательных подбрасываний монеты, приводит к точно такому же вероятностному распределению, называемому нормальным распределением (или колоколообразной кривой). Доска Гальтона — просто наглядное, зримое выражение теоремы Лапласа.

Центральная предельная теорема — воистину чудо математики *XIX* века. Только задумайтесь: хотя путь каждого отдель-

ного шарика непредсказуем, путь тысячи шариков может быть предсказан довольно точно — удобный факт для продюсеров викторины «Цена верна», которые могут подсчитать, сколько денег все участники выиграют за отчетный период. Этот же закон нужно благодарить за то, что страхование от несчастных случаев оказывается весьма надежным и прибыльным делом, хотя пути Господни для отдельной человеческой судьбы неисповедимы.

Хорошо одетая публика в Королевском институте, вероятно, недоумевала: какое всё это имеет отношение к законам наследуемости — заявленной теме доклада? Чтобы продемонстрировать связь, Гальтон представил слушателям данные, полученные во Франции, где измерялся рост солдат-призывников. У этого показателя распределение тоже оказалось нормальным: людей с ростом около среднего больше всего, а в обе стороны от среднего, по направлению к самым высоким и самым низким, их число плавно уменьшается. На самом деле неважно, о чем идет речь, о росте тысячи призывников или о тысяче шариков в пазах доски Гальтона, если число категорий в выборке (пазов или ростовых промежутков) будет одинаковым, то сравнительно одинаковым будет и распределение индивидуальных случаев по категориям от центра до краев.

Таким образом, по Гальтону, его прибор представляет собой модель наследования роста, как, впрочем, и многих других наследственно обусловленных признаков. Это каузальная модель. Иными словами, согласно Гальтону, каждый шарик «наследует» свое положение на доске примерно по такому же механизму, по которому люди наследуют рост.

Но если мы принимаем эту модель — временно, — то обнаруживается загадка, о которой Гальтон и собирался рассказать тем вечером. Ширина колоколообразной кривой зависит от числа рядов булавок, расположенных между верхней и нижней стороной доски. Допустим, мы удвоим число рядов. Это будет моделью наследования в двух поколениях, первая половина рядов будет соответствовать первому поколению, а вторая — второму. В этом случае мы неизбежно обнаружим большее разнообразие вариантов значений во втором поколении

по сравнению с первым, и с каждым последующим поколением колоколообразная кривая будет становиться все шире и шире.

Однако с ростом человека ничего подобного не происходит. Ширина распределения роста людей остается более-менее постоянной с течением времени. Людей трехметрового роста не встречалось 100 лет назад, нет их и сейчас. Что обусловливает стабильность подобных признаков в популяции? Гальтон размышлял над этой загадкой примерно восемь лет, с момента выхода его сочинения «Наследственный гений» в 1869 году.

Как и предполагает заглавие книги, на самом деле Гальтона интересовали не детские настольные игры и не рост солдат, а наследование интеллектуальных способностей человека. Будучи представителем большого круга родства, из которого вышло много выдающихся ученых, Гальтон вполне ожидаемо хотел бы показать, что талант — свойство фамильное, и именно этому он и посвятил свою книгу. Он дотошно составил родословные 605 «выдающихся» англичан, живших в течение четырех предшествующих столетий. Однако обнаружилось, что сыновья этих замечательных граждан, равно как и отцы, были заметно менее исключительными, а их деды и правнуки — еще малопримечательнее.

Сейчас нам нетрудно найти недостатки в постановке задачи, предложенной Гальтоном. Во-первых, возможно ли дать точное определение, что такое «выдающесть»? И не окажется ли, что люди из выдающихся семейств успешны благодаря доступным им привилегиям, а не благодаря таланту? Хотя Гальтон и осознавал эти сложности, он продолжал свои бесплодные поиски генетического определения таланта со все возрастающим рвением.

Тем не менее ученый обнаружил кое-что весьма занимательное, что стало еще более очевидным, когда он переключился на такие признаки, как рост, который проще измерить и который связан с наследственностью более явно, чем талант. Сыновья высоких мужчин, как правило, выше среднего роста, хотя и не такие высокие, как их отцы. Гальтон назвал это явление сначала реверсией, а потом регрессией к среднему значению. Это же явление наблюдается во многих других ситуациях. Если

школьники выполняют две разные, но стандартизованные контрольные работы по одному и тому же материалу, то те, кто имел самые высокие баллы за первую контрольную, получат оценки выше среднего и за вторую, хотя и не такие высокие, как в первый раз. Феномен возвращения к среднему встречается повсеместно во всех сферах жизни, образования и бизнеса. Так, в бейсболе новичок года, показавший неожиданно высокие результаты, на втором году обычно «провисает» и играет уже не так хорошо.

Гальтон не знал подобных примеров и предполагал, что наткнулся на закон наследования, а не на закон статистики. Он полагал, что возвращение к среднему обусловлено некой причиной, и на лекции в Королевском институте наглядно проиллюстрировал свои доводы, представив публике двухуровневый квинкункс.

Пройдя первый ряд булавок, шарики попадали в наклонные пазы, которые смещали их вновь к центру доски; затем они проходили второй ряд. Гальтон торжественно показал, что эти пазы полностью компенсируют тенденцию нормального распределения расползаться вширь. В этом случае колоколообразная кривая распределения вероятностей оставалась одной и той же ширины от поколения к поколению.

Таким образом, постулировал Гальтон, возвращение к среднему — это физический процесс, с помощью которого природа обеспечивает одинаковое распределение роста (или интеллекта) в каждом последующем поколении. «Процесс регрессии сотрудничает с общим законом отклонения», — сообщил он своей аудитории. Ученый сравнил его с законом Гука, описывающим тенденцию пружины возвращаться к равновесной длине.

Не забываем, какой был год на дворе. В 1877 году Гальтон искал причинное объяснение и полагал, что регрессия к среднему — это каузальный процесс, подобный закону физики. Он ошибался, но был в этом не одинок. Многие повторяют эту ошибку по сей день. Например, бейсбольные эксперты почти всегда пытаются объяснить «проседание» чемпиона на втором году рассуждениями о причинах. «Он зазнался и расслабился», — сетуют они, или: «Другие игроки сумели найти его

слабости и воспользоваться ими». Это может быть правдой, но на деле такой феномен не нуждается в объяснении причин. Чтобы оно произошло, обычно достаточно просто закона случая.

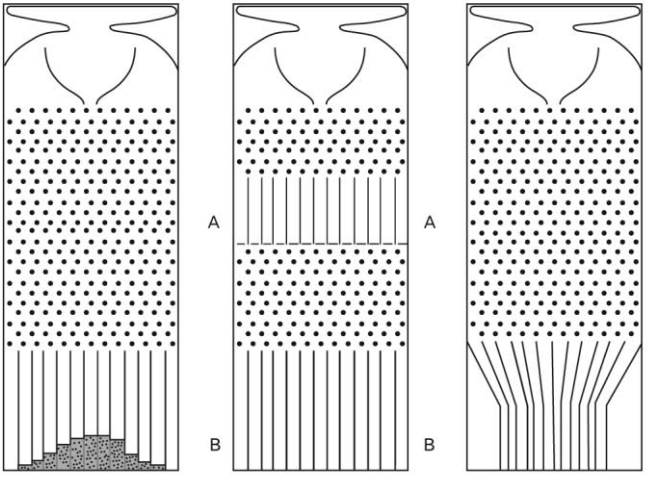


Рис. 8. Устройство, с помощью которого Фрэнсис Гальтон продемонстрировал аналогию наследования роста: *а* — когда через устройство вроде пинбола пропускают большое количество шариков, то в результате случайных отскакиваний они накапливаются на дне устройства, образуя колоколообразную кривую; *б* — при двух проходах через устройство, соответствующих двум поколениям, кривая распределения становится шире; *в*-чтобы упредить эту тенденцию, Гальтон придумал желобки, возвращающие шарики к центру во «втором поколении». Этими желобками Гальтон каузально объясняет явление возвращения к среднему [источник: Гальтон Ф. Естественная наследственность (1889)]

Современная статистика объясняет это явление совсем просто. В книге «Думай медленно, решай быстро» Даниэль Каneман делает вывод: «Успех — это талант плюс удача. Большой успех — это чуть больше таланта и намного больше удачи». Новичок года талантливее в бейсболе, чем большинство, но ему, скорее всего, еще и очень повезло. В следующем году ему повезет меньше и его баллы окажутся не столь впечатляющими.

К 1899 году Гальтон это понял и в процессе постижения, разочарованный, но одновременно и восхищенный открывающимся, предпринял первый значительный шаг к отделению статистического от причинного. Его рассуждения несколько туманны, но их стоит попытаться понять — ведь это первый, пока робкий лепет только что родившейся статистики.

Гальтон стал собирать разнообразные, так называемые антропометрические данные: рост, длину предплечья, длину и ширину головы и т.п. Он заметил, что если два размерных признака, например рост и длину предплечья, расположить на оси координат, то их сочетание проявляет все ту же регрессию к среднему. У самых высоких людей более длинные руки, чем в среднем, но длина их рук не настолько больше среднего, насколько рост. При этом очевидно, что рост не является причиной длины руки или, наоборот, в лучшем случае и то и другое имеют общую наследственную компоненту. Гальтон стал использовать новый термин для таких пар признаков: рост и длина предплечья со-отнесены, находятся в ко-реляции, ко-релируют. Со временем он перешел к более привычному нам написанию: «корреляция», «коррелируют».

Чуть позже он обнаружил еще более неожиданный факт: при сравнении поколений неважно, движемся ли мы по ходу времени или назад в прошлое. Это значит, что отцы относительно сыновей тоже проявляют возвращение к среднему. Отец сына, который выше ростом, чем популяция в среднем, оказывается почти всегда тоже выше среднего роста, но ниже, чем его сын (рис. 9). Заметив это, Гальтон был вынужден отказаться от попыток найти каузальное объяснение явлению регрессии к среднему, потому что рост сына никоим образом не может определять рост отца.

На первый взгляд, это наблюдение парадоксально. «Постойте! — скажете вы. — Значит, у более длинных отцов более короткие сыновья, а у более длинных сыновей более короткие отцы? Как эти два утверждения могут быть верны одновременно? Не может же сын быть одновременно выше и ниже своего отца».

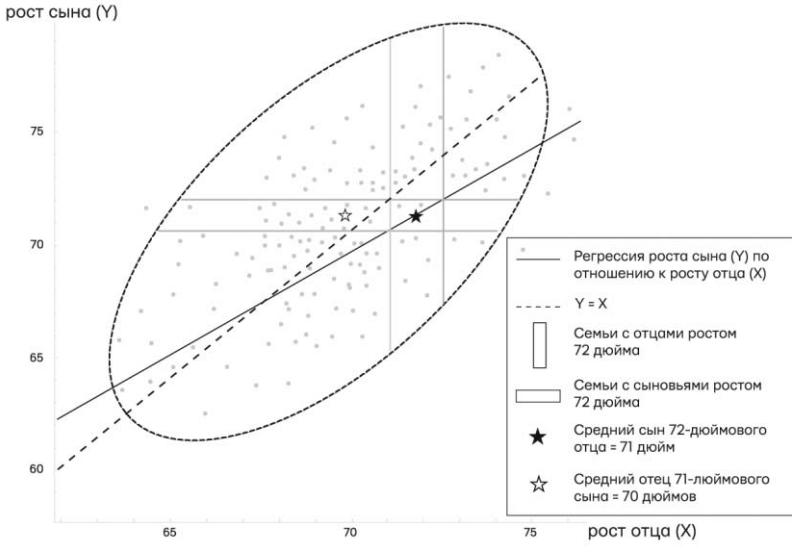


Рис. 9. Точечный график показывает набор данных о росте. Каждая точка представляет рост отца (по оси X) и сына (по оси *Y*). Пунктирная линия совпадает с большой осью эллипса, а сплошная линия (ее называют линией регрессии) соединяет крайнюю левую и крайнюю правую точки эллипса. Разница между ними отражает возвращение к среднему. Например, черная звездочка показывает, что у отцов ростом 72 дюйма сыновья в среднем имеют рост 71 дюйм (т.е. средний рост у всех, чьи данные представлены точками в вертикальной полосе,— 71 дюйм). Горизонтальная полоса и белая звездочка показывают, что такое же падение роста возникает в непричинном направлении (назад во времени) (источник: график Маян Харел при участии Кристофера Баучера)

Ответ заключается в том, что мы говорим не об индивидуальных отцах и сыновьях, а о двух популяциях — отцовской и сыновней. Допустим, мы отобрали отцов, чей рост ровно 6 футов. Это больше среднего, поэтому средний рост их сыновей будет тоже выше среднего, но ближе к среднему, допустим, 5 футов и 11 дюймов. Однако множество пар, в которых рост отца равен 6 футам, не совпадает с множеством пар, в которых рост сына — 5 футов 11 дюймам. В первом множестве рост

всех отцов равен 6 футам по условию задачи, а вот во втором окажется несколько отцов с ростом больше 6 футов и много отцов ниже 6 футов. Их средний рост будет ниже, чем 5 футов 11 дюймов, и таким образом регрессия к среднему снова обнаружит себя.

Другой способ наглядно изобразить регрессию — построить диаграмму, называемую точечным графиком (см. рис. 9). Каждая пара из отца и сына на нем представлена точкой, при этом ее положение по оси *X* определяется ростом отца, а по оси *Y* — ростом сына. Таким образом, отец и сын, оба ростом 5 футов 9 дюймов (или 69 дюймов), вместе окажутся на графике точкой с координатами (69; 69) прямо по центру точечного графика. Отец ростом 6 футов (или 72 дюйма) и сын ростом 5 футов 11 дюймов (71 дюйм) попадут в точку (72; 71) в северо-западной части нашей диаграммы. Обратите внимание, что облако полученных точек приближается по форме к эллипсу — факт, принципиальный для анализа Гальтона и характерный для нормального распределения для двух признаков.

Как показано на рис. 9, пары, в которых отцы ростом 72 дюйма, располагаются в вертикальном сегменте эллипса с центром в точке 72, а пары, в которых рост сыновей 71 дюйм, расположены в горизонтальном сегменте с центром в точке 71, что графически доказывает, что это две разные выборки. Сосредоточившись только на первой из них, парах с отцами ростом 72 дюйма, мы зададим вопрос, каков средний рост сыновей или, что то же самое, где находится центр этого вертикального сегмента (на глаз можно прикинуть, что центр приходится примерно на 71). Если мы рассмотрим только вторую выборку, в которой рост сыновей 71 дюйм, и спросим, каков средний рост их отцов, это будет равносильно нахождению центра горизонтального сегмента — легко увидеть, что он находится где-то на отметке 70,3.

Двигаясь дальше, выполняем такую же процедуру для всех вертикальных сегментов. Это равносильно вопросу «Каков наиболее вероятный рост сыновей (*Y*) для отцов ростом *X*?». И наоборот, рассматривая все горизонтальные сегменты, выясняем, где центр каждого из них: каким окажется (вернее,

был, тут мы предсказываем прошлое) наиболее вероятный рост отцов для сыновей с ростом *Y*.

Размышляя над этими вопросами, Гальтон подошел к важному моменту: предсказания всегда располагаются на линии, названной им линией регрессии, которая расположена более полого, чем главная ось (или ось симметрии) данного эллипса. На самом деле таких линий две — в зависимости от того, данные каких из двух переменных известны и взяты в качестве основания для прогноза, а какие надо предсказать. Можно предугадать рост сыновей по росту отцов, а можно и наоборот. Ситуация совершенно симметрична. И это еще раз демонстрирует нам, что в случаях, где наблюдается регрессия к среднему, между причиной и следствием нет разницы.

Наклон линии регрессии позволяет нам предсказывать значение одной переменной, если нам известны значения второй. В терминах задачи Гальтона наклон в 0,5 означает, что каждому дюйму сверх среднего в росте отца соответствуют дополнительные полдюйма роста сына и наоборот. Наклон, равный единице, свидетельствовал бы о точной корреляции, т.е. каждый дополнительный дюйм роста у отца передавался бы по наследству сыну, который тоже был бы на этот дюйм выше. Наклон кривой не бывает больше единицы: в таком случае сыновья высоких отцов были бы в среднем выше, а сыновья отцов небольшого роста были бы ниже последних, а распределение роста в популяции становилось бы со временем все шире и шире. Через несколько поколений некоторые люди были бы трехметрового роста, а другие — ростом меньше метра, чего в природе не наблюдается. Таким образом, если распределение признака остается одинаковым от поколения к поколению, наклон линии регрессии не превышает единицы.

Закон регрессии применим даже тогда, когда мы рассматриваем корреляцию двух совсем разных признаков, например рост и ай-кью. Если расположить значения одного признака относительно значений другого на точечном графике и правильно подобрать масштаб обеих осей, наклон наиболее близко подходящей прямой всегда будет обладать теми же свойствами. Он равен единице только тогда, когда значения одного

признака можно четко предсказать по значениям другого; он равен нулю, если связи между признаками нет и предсказание равносильно случайности. После масштабирования наклон прямой одинаков вне зависимости от того, рассматриваем ли мы признак X относительно признака *Y* или наоборот. Другими словами, наклон прямой ничего не говорит нам о том, что в данном случае причина, а что следствие. Одна переменная обусловливает значения другой, или обе они обусловливаются третьей; для предсказания их значений это не важно.

Гальтонова идея корреляции впервые предоставила объективную меру связи двух переменных друг с другом, не зависящую от человеческих суждений и интерпретаций. Эти две переменные могут быть ростом, интеллектом или уровнем дохода; они могут находиться в каузальной, нейтральной или обратно-каузальной зависимости друг от друга — их корреляция всегда будет отражать степень взаимной предсказуемости значений двух признаков. Ученик Гальтона Карл Пирсон позже вывел формулу для наклона (правильно масштабированной) линии регрессии и назвал ее коэффициентом корреляции. До сих пор это первое число, которое вычисляют статистики по всему земному шару, когда хотят узнать, насколько взаимосвязаны любые два признака в массиве данных. Гальтон и Пирсон, должно быть, пришли в восторг, обнаружив такой универсальный способ описания взаимоотношений между случайными переменными. Старые, скользкие концепции причины и следствия по сравнению с математически прозрачной и четкой концепцией коэффициента корреляции казались устаревшими и ненаучными, в особенности Пирсону.

Гальтон и оставленные поиски

По иронии истории Гальтон начал с поисков причинности, а закончил открытием корреляции, отношения, лишенного причинности. Однако все равно признаки каузального мышления остаются в его публикациях. «Легко заметить, что корреляция [между размерами двух органов] должна быть следствием того,

что изменчивость двух этих органов отчасти вызвана общими причинами», — пишет он в 1889 году. Первым жертвоприношением на алтарь корреляции стала сложная машина Гальтона для объяснения стабильности распределения генетических признаков в популяции. Доска Гальтона имитировала создание изменчивости по длине тела и ее передачу от поколения к поколению. Но ученому пришлось изобрести наклонные желоба в своей машине, ограничивающие постоянно возрастающее разнообразие в популяции. Не сумев обнаружить биологический механизм, удовлетворительно объясняющий эту силу, возвращающую к среднему, Гальтон просто прервал попытки после восьми лет бесплодных поисков и все внимание сосредоточил на корреляции, как моряк на песне сирены. Статистик Стивен Стиглер, много писавший о Гальтоне, заметил этот неожиданный сдвиг в целях и ожиданиях ученого: «Фигурой умолчания оказались Дарвин, желобки, все это „выживание наиболее приспособленных". … По жестокой иронии, то, что начиналось как попытка подвести математическую основу под „Происхождение видов", закончилось тем, что сама суть этой великой работы оказалась отброшена, как ненужная!»

Но для нас, живущих в современную эпоху причинного вывода, исходная проблема остается. Как мы объясним стабильность популяционного среднего, невзирая на дарвиновскую изменчивость, которой одно поколение наделяет последующее?

Возвращаясь к машине Гальтона в свете диаграмм причинности, первое, что я замечаю, — это то, что она была сконструирована неправильно. Постоянно растущей дисперсии, которая вынудила ученого создать ей противовес, вообще не должно было там быть. В самом деле, если мы проследим падение шарика в доске Гальтона с одного уровня на другой, мы увидим, что отклонение на следующем уровне наследует сумму всех отклонений, причиненных всеми булавками, с которыми он сталкивался на своем пути. Это откровенно противоречит уравнению Канемана:

Успех = Талант + Удача;

Большой успех = Чуть больше таланта + Намного больше удачи.

Согласно этим уравнениям, успех в поколении 2 не наследует удачу из поколения 1. Удача по определению преходяща и случайна; она не может влиять на будущие поколения. Но подобное поведение признака несовместимо с устройством машины Гальтона. Чтобы сравнить эти две концепции рядом, нарисуем их ассоциированные диаграммы причинности. На рис. 10а (концепция Гальтона) успех передается через поколения и удача накапливается неограниченно. Это легко себе представить, если под успехом понимать богатство или знатность. Однако для описания наследования физических характеристик, таких как рост, нам придется заменить модель Гальтона той, что на рис. 10б. В ней только генетическая компонента, показанная здесь как талант, передается от одного поколения к другому. Удача действует на каждое поколение независимо, таким образом, что случайные факторы в одном поколении не могут влиять на последующие поколения ни прямо, ни косвенно.

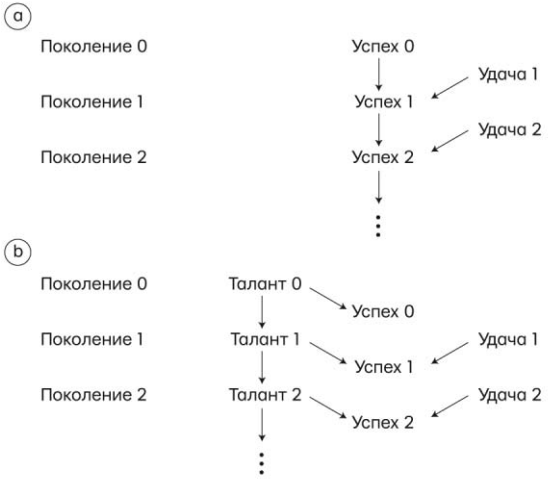


Рис. 10. Две модели наследуемости: *а* — модель, соответствующая машине Гальтона, в которой удача накапливается от поколения к поколению, приводя ко все возрастающей дисперсии успеха; *б* — генетическая модель, в которой удача не накапливается, приводит к постоянному разбросу успеха

Обе эти модели совместимы с колоколообразным распределением значений роста. Но первая модель не совместима со стабильностью разброса роста (или успеха). Вторая же модель показывает, что для объяснения стабильности разброса успеха от поколения к поколению нам достаточно объяснить только стабильность генетических факторов в популяции (таланта). Эта стабильность, теперь называемая равновесием Харди — Вайнберга, получила удовлетворительное математическое объяснение в работе Годфри Харолда Харди и Вильгельма Вайнберга 1908 года. И да, они основывались на еще одной каузальной модели — менделевской теории наследственности.

Ретроспективно рассуждая, Гальтон не мог предвидеть достижения Менделя, Харди и Вайнберга. В 1877 году, когда Гальтон прочитал свою лекцию, работа Грегора Менделя 1866 года была основательно забыта (ее вновь открыли только в 1900 году), а математические выкладки доказательства Харди и Вайнберга были бы для него, вероятно, слишком сложны. Однако интересно обратить внимание, как близок он был к верному подходу и как диаграммы причинности легко вскрывают ложность его допущения: передачу случайных факторов, удачи, от одного поколения к другому. К сожалению, его завела в тупик красивая, но неверная причинная модель, а позже, открыв красоту корреляции, он уже решил, что каузальность больше не нужна.

В качестве заключительного личного комментария к истории Гальтона я сознаюсь в смертном для историка грехе, одном из многих грехов, допущенных мной в этой книге. В 1960-х стало немодно писать историю науки с точки зрения современной науки, как я делал выше. Такой стиль исторических заметок, который фокусируется на удачных теориях и экспериментах и уделяет мало внимания неподтвержденным теориям и научным тупикам, теперь именуют издевательским термином «история в стиле вигов». Современный стиль истории науки более демократичен, в нем химики и алхимики пользуются равным уважением, а все теории рассматриваются в социальном контексте своего времени.

Однако, когда приходится объяснять, каким образом причинность была изгнана из статистики, я с гордостью надеваю

парик историка-вига. Иначе как надев каузальные очки и пересказав историю Гальтона и Пирсона в свете современной науки о причинах и следствиях, просто невозможно понять, как же статистика стала нечувствительным к типу модели методом редукции данных. На самом деле, поступая так, я выпрямляю искажения, созданные современным большинством историков, которые, не владея каузальным словарем, восхищаются изобретением корреляции и не способны заметить огромную потерю — смерть причинности.

Гнев фанатика

Завершить позорное изгнание причинности из статистики предстояло ученику Гальтона Карлу Пирсону. Однако даже он не смог довести это изгнание до конца.

Прочтение «Естественной наследственности» Гальтона стало одним из определяющих моментов в жизни Пирсона: «Я почувствовал себя корсаром времен Дрейка — членом отряда „не вполне пиратов, но с отчетливо пиратскими тенденциями", как уточняет словарь! — написал он в 1934 году. — Я понял… Гальтона так, что он предполагал существование категории более широкой, чем причинная связь, а именно корреляции, по отношению к которой причинность была только предельным состоянием, и эта новая концепция корреляции ввела психологию, антропологию, медицину и социологию в значительной степени в поле математического анализа. Именно Гальтон впервые освободил меня от предрассудка, будто достойная математика может быть применима только к природным явлениям под категорией причинной связи».

Согласно взглядам Пирсона, Гальтон расширил словарь науки. Причинно-следственные связи были разжалованы в не более чем частный случай корреляции (а именно когда коэффициент корреляции равен 1 или –1 и взаимоотношения между X и Y жестко детерминированы). Свое видение причинности он очень четко формулирует в «Грамматике науки» (1892): «То, что определенная последовательность явлений случалась

и повторялась в прошлом, — это опыт, который мы выражаем в концепции причинно-следственных связей. … Наука никоим образом не может продемонстрировать никакой неотъемлемой необходимости в последовательности явлений, ни доказать с абсолютной уверенностью, что эта последовательность должна воспроизводиться». Таким образом, причинность для Пирсона сводится к повторяемости и в детерминистском смысле не может быть доказана. К причинности в недетерминистском мире Пирсон еще более пренебрежителен: «В конечном итоге научное утверждение, описывающее отношение двух явлений, всегда может быть сведено… к таблице сопряженности».

Другими словами, наука — это только данные. Больше ничего. В этом мировоззрении понятия действия и альтернативного сценария, обсужденные в главе 1, не существуют, и самый нижний уровень Лестницы Причинности — это все, что нужно, чтобы заниматься наукой.

Ментальное сальто от Гальтона к Пирсону захватывает дух и действительно достойно корсара. Гальтон доказал только, что одно явление — регрессия к среднему — не нуждается в каузальном объяснении. Пирсон же полностью удалил причинность из науки. Что привело его к этому логическому трюку?

Историк Тед Портер в написанной им биографии «Карл Пирсон» рассказывает, что скептицизм по отношению к причинности был у Пирсона и до прочтения книги Гальтона. Пирсон боролся с философскими основаниями физики и писал, например: «Полагать силы причиной движения так же обоснованно, как думать, что рост дерева вызывают дриады». С более общей точки зрения Пирсон принадлежал к течению, именуемому позитивизмом, согласно которому Вселенная — это производная человеческой мысли, а наука — только описание этой мысли. Таким образом, причинность, понимаемая как объективный процесс, происходящий в мире снаружи человеческого мозга, не могла иметь в этой концепции никакого научного значения. Значащие мысли способны только отражать наблюдения, а последние полностью описываются с помощью корреляций. Решив, что корреляция гораздо более универсально описывает

человеческое мышление, чем причинность, Пирсон приготовился к тому, чтобы избавиться от причинности окончательно.

Портер рисует яркий, живой портрет Пирсона, всю жизнь называвшего себя немецким словом *SchWärmer*, которое обычно переводится как «энтузиаст», но может иметь и более резкое значение — «фанатик». Окончив Кембридж в 1879 году, Пирсон провел год в Германии и так полюбил немецкую культуру, что изменил первую букву своего имени Карл (*Carl*), с *C* на *K*, на немецкий манер. Задолго до того, как это стало модно, он придерживался социалистических взглядов, и в 1881 году написал Карлу Марксу, предлагая перевести «Капитал» на английский. Пирсон, по некоторым мнениям первый английский феминист, основал лондонский «Клуб мужчин и женщин» для обсуждения «женского вопроса». Его волновал низкий статус женщин в обществе, и он настаивал на том, чтобы им достойно платили за работу. К идеям он относился с большой страстью — и одновременно очень рассудочно к своим страстям. Ему понадобилось почти полгода, чтобы уговорить свою будущую жену Марию Шарп выйти за него, и из их переписки понятно, что она откровенно опасалась, что не сможет соответствовать его высоким интеллектуальным идеалам. Когда Пирсон открыл для себя Гальтона и его корреляции, его страстность наконец-то нашла точку приложения; эта идея, как он полагал, могла перевернуть мир науки и привнести математическую строгость в такие области, как биология и психология. К достижению этой цели он и ринулся с поистине пиратской целеустремленностью. Его первая статья о статистике вышла в 1893 году, через четыре года после открытия корреляции Гальтоном. В 1901 году он основал журнал «Биометрика» (*Biometrika*), до наших дней остающийся одним из самых влиятельных статистических журналов (в нем была еретически опубликована моя первая статья по диаграммам причинности в 1995 году).

К 1903 году Пирсон получил грант от Почетной компании драпировщиков на создание лаборатории биометрии в Университетском колледже Лондона. В 1911 году она стала официальным факультетом, когда Гальтон умер и оставил средства на создание профессорской кафедры (с условием,

что первым профессором на ней станет Пирсон). По крайней мере два десятилетия пирсоновская лаборатория биометрии была ведущим мировым статистическим центром. Когда Пирсон получил руководящую должность, его фанатизм стал проявляться все более выраженно. Вот что пишет Портер: «Возглавляемое Пирсоном статистическое движение имело все признаки раскольнической секты. От своих соратников он требовал лояльности и самоотверженности, а оппонентов отлучал от церкви биометрии». Один из его первых ассистентов Джордж Юл оказался также одним из первых, на кого обрушился его гнев. Некролог Пирсону, написанный Юлом для Королевского общества в 1936 году, хорошо передает тогдашнюю злобу дня, хотя и написан сдержанно, огибая острые углы: «Заразительность его энтузиазма была действительно бесценна; но доминирование, даже в готовности помочь, было несомненным недостатком. … Это страстное желание доминировать, чтобы все было именно так, как ему хочется, проявлялось и во многом другом, например в редактировании „Биометрики" — ни один журнал в мире не редактировался с таким личным пристрастием. … Те, кто оставил его и начал мыслить самостоятельно, обнаруживали, один за другим, что после расхождения мнений поддерживать дружеские отношения с ним оказывалось крайне сложно, а после прямой критики — невозможно».

Тем не менее в возведенной Пирсоном оборонной башне науки без причинности находились трещины, причем чаще по вине его соратников-основателей, чем поздних учеников. Так, сам Пирсон неожиданно написал несколько статей о «ложных корреляциях», о понятии, которое невозможно ввести без отсылок к причинности. Пирсон заметил, что довольно легко найти корреляции, которые просто очевидно бессмысленны. В качестве забавного примера в постпирсоновские времена часто приводили тот факт, что существует высокая корреляция между потреблением шоколада на душу населения в странах мира и числом нобелевских лауреатов в этих же странах. Эта корреляция выглядит глупо, потому что нельзя вообразить, каким образом шоколад на десерт может сделать человека

нобелевским лауреатом. Правдоподобное объяснение заключается в предположении, что в преуспевающих странах Запада люди могут позволить себе больше шоколада, а премию Нобеля получают также в основном выходцы из этих наиболее развитых стран. Но это типичное каузальное объяснение, которое, согласно Пирсону, не требуется для научного мышления. Для него причинность — только «фетиш в непостижимой магии современной науки». Корреляция должна быть целью научного понимания. Этот подход, однако, ставит его в неловкое положение, когда ему приходится объяснять, почему одни корреляции имеют смысл, а другие «ложны». Он поясняет, что истинная корреляция указывает на «органическую связь» между переменными, в то время как для ложной корреляции такой связи нет. Но что такое органическая связь? Разве это не та же причинность, только под другим именем?

Вместе Пирсон и Юл собрали несколько случаев ложных корреляций. Одна их категория теперь называется смешением, история с нобелевскими лауреатами и шоколадом — типичный ее образец (уровень благосостояния и местоположение — смешанные факторы, или общие причины для уровня потребления шоколада и числа лауреатов премии Нобеля). Другой пример бессмысленной корреляции часто обнаруживается при анализе серий данных, изменяющихся во времени. Так, Юл нашел невероятно высокую корреляцию (0,95) между уровнем смертности в Англии в данный год и процентом браков, заключенных в тот же год в англиканской церкви. Неужели Бог избирательно наказывает сочетающихся законным браком англикан? Конечно, нет! Две совершенно отдельных исторических тенденции просто совпали по времени: смертность в стране неуклонно сокращалась, а число членов англиканской церкви так же неуклонно уменьшалось. Поскольку оба процесса шли в одном направлении в одно и то же время, между ними была положительная корреляция при отсутствии причинной связи.

Самый интересный вариант бессмысленной корреляции Пирсон обнаружил еще в 1899 году. Он проявляется тогда, когда две гетерогенные выборки объединяют в одну. Пирсон, который, как и Гальтон, фанатично собирал данные, относя-

щиеся к человеческому телу, получил обмеры 806 мужских и 340 женских черепов из парижских катакомб и подсчитал корреляции между длиной и шириной черепа. Когда подсчет производился только для мужских или только для женских черепов, корреляция была пренебрежимо мала — между длиной и шириной черепа практически не было связи. Но если обе группы объединяли, корреляция становилась равной 0,197, и обычно такое значение считалось значимым. Это объяснимо, потому что небольшая длина черепа сегодня считается индикатором того, что череп принадлежал женщине, и поэтому его ширина тоже окажется небольшой. Тем не менее Пирсон считал это статистическим артефактом.

Тот факт, что корреляция оказалась положительной, не имел биологического или «органического» значения; это был просто результат неправомерного объединения двух разных выборок.

Этот пример являет собой частный случай более общего явления, именуемого парадоксом Симпсона. В главе 6 мы обсудим, в каких случаях оправдано разделение массива данных на отдельные группы, и объясним, почему при их объединении могут возникать ложные корреляции.

Но давайте взглянем на то, что писал Пирсон: «Для тех, кто настаивает на взглядах на любые корреляции как на связь причины и следствия, тот факт, что значимую корреляцию между двумя совершенно не связанными признаками А и Б можно получить искусственным смешением двух близких выборок, должен восприниматься как шок». Стивен Стиглер комментирует это: «Я не могу удержаться от догадки, что сильнее всего был шокирован он сам». По сути, Пирсон бранил сам себя за склонность мыслить в терминах причинности.

Глядя на этот же самый пример через линзу причинности, нам остается только воскликнуть: надо же было упустить такую возможность! В идеальном мире подобные случаи могли бы подвигнуть талантливого ученого на размышления о причинах его шока и разработку научной дисциплины, предсказывающей появление ложных корреляций. По крайней мере, он попытался бы объяснить, когда данные целесообразно объединять, а когда нет. Но единственное наставление Пирсона последователям

по этому поводу заключается в том, что «искусственное» (что бы это ни значило) объединение данных — это плохо. По иронии судьбы, используя наши каузальные очки, мы теперь знаем, что иногда именно анализ объединенных, а не разделенных данных дает верный ответ. Логика причинных умозаключений может подсказать нам, чему следует доверять. Я бы хотел, чтобы Пирсон был сейчас с нами и мог этому порадоваться!

Далеко не все ученики Пирсона ступали за ним след в след. Юл, который разошелся с Пирсоном по другим причинам, по этому поводу тоже был с ним не согласен. Вначале он был с ним в одном экстремистском лагере, утверждая, что корреляции расскажут нам все, что мы могли бы захотеть узнать посредством науки. Тем не менее он до некоторой степени изменил свое мнение, когда ему понадобилось объяснить наблюдения за условиями жизни беднейших жителей Лондона. В 1899 году он изучал вопрос, увеличивает ли «внешняя помощь» (материальная помощь, доставляемая на дом малоимущим, в отличие от жизни в богадельне) уровень бедности. Данные показывали, что кварталы, получающие больше «внешней помощи», отличались более высоким уровнем бедности, но Юл понял, что эта корреляция, скорее всего, была ложной; в этих кварталах жило больше пожилых людей, которые чаще всего бедны. Однако затем он сумел показать, что при сравнении кварталов с одинаковой пропорцией пожилых жителей корреляция сохраняется. Благодаря этому он осмелился заявить, что повышение уровня бедности действительно связано с «внешней помощью». Однако, выйдя из строя, чтобы сделать это утверждение, он поспешил вернуться в строй, написав в примечании: «Строго говоря, „по причине" следует читать как „связано с"». Целые поколения ученых после него следовали этому образцу. Они думали: «А происходит по причине Б», но говорили: «А связано с Б». Однако Пирсон с последователями, активно выступающие против причинности, и колеблющиеся недодиссиденты вроде Юла, опасающиеся разозлить лидера, подготовили сцену к выступлению нового игрока — ученого из-за океана, который бросил первый откровенный вызов научной культуре, избегающей понятия причинности.

Сьюалл Райт, морские свинки и путевые диаграммы

Когда Сьюалл Райт прибыл в Гарвардский университет в 1912 году, его образование на тот момент вряд ли предсказывало долговременный эффект, который он произведет в науке. Он учился в маленьком (и ныне закрытом) колледже Ломбард в Иллинойсе, и в его выпуске было всего семь студентов. Одним из преподавателей был его собственный отец Филип Райт — швец, жнец и на дуде игрец от науки, на нем держалась даже типография колледжа. Сьюалл и его брат Квинси помогали отцу в печатном деле, и помимо прочего в их типографии был издан первый сборник тогда еще неизвестного поэта и студента Ломбарда Карла Сэндберга.

Сьюалл Райт поддерживал тесную связь с отцом еще долгие годы после окончания колледжа. Когда Сьюалл переехал в Массачусетс, папа Филип последовал за ним. Позже, когда Сьюалл работал в Вашингтоне, там же трудился и Филип, сначала в Американской тарифной комиссии, а потом в Брукингском институте экономистом. Хотя их академические интересы сильно разошлись, они находили способы сотрудничать, и Филип стал первым экономистом, использовавшим путевые диаграммы, изобретенные его сыном.

Райт-младший приехал в Гарвард изучать генетику, в то время одно из самых активно развивавшихся направлений в науке, потому что теория Грегора Менделя о доминантных и рецессивных генах была только что открыта заново. Научный руководитель Райта Уильям Касл идентифицировал восемь различных наследственных факторов (или генов, как бы мы назвали их сегодня), влияющих на цвет меха у кроликов. Касл предложил Райту провести аналогичное исследование на морских свинках. Защитив диссертацию в 1915 году, Райт получил предложение работы, для которой никто не подходил лучше него: работать с морскими свинками в Департаменте сельского хозяйства США (*United States Department of Agriculture; USDA*).

Сейчас остается только гадать, понимали ли в департаменте, кого они берут на работу. Возможно, им просто нужен был ответственный зоотехник, который мог бы привести в порядок 20-летний архив, все это время представлявший собой полный хаос. Райт сделал не только это, но и намного, намного больше. Морские свинки для Райта стали движущей пружиной всей его карьеры и ключевым звеном в его теории эволюции, совсем как галапагосские вьюрки для Дарвина. Райт был одним из ранних приверженцев идеи, что эволюция не идет постепенно, как предполагал Дарвин, а происходит посредством относительно внезапных рывков.

В 1925 году Райт перешел на ставку на кафедре в Чикагском университете, которая, вероятно, лучше подходила человеку столь разносторонних научных интересов. Однако и там он по-прежнему оставался предан морским свинкам. Часто рассказывают анекдот, что однажды во время лекции он держал под мышкой особо буйную морскую свинку и по рассеянности вдруг начал стирать ей с доски вместо тряпки. Хотя все жизнеописатели Райта согласны, что эта история скорее всего апокриф, подобные детали обычно говорят о личности намного больше, чем сухие биографии.

Нас в этой главе больше всего интересует начало работы Райта в *U*SDA. Наследование окраски меха у морских свинок упорно отказывалось подчиняться законам Менделя. Оказалось практически невозможным получить чисто белую или разноцветную свинку, и даже самые инбредные линии (после многих поколений скрещиваний только между братьями и сестрами) все еще обнаруживали значительную изменчивость окраски, от преимущественно белой до преимущественно разноцветной. Это противоречило предсказанию менделевской генетики, согласно которому после большого числа поколений близкородственных скрещиваний признак «закрепляется». Райт начал сомневаться, что процент белого в окраске определяется одной только генетикой, и постулировал, что часть изменчивости определяется «внутриутробными факторами» во время беременности. Задним числом мы знаем, что он был прав. Различные гены окраски экспрессируются по-разному в различных частях

тела, и распределение окраски зависит не только от генов, которые унаследовало животное, но и от того, где именно и в каких комбинациях будет происходить их экспрессия или подавление.

Как это часто случается (по крайней мере, с гениями!) под давлением требующей решения проблемы на свет появился новый метод анализа, который теперь применяется гораздо шире, чем в родной генетике морских свинок. Однако для Сьюалла Райта оценка внутриутробных факторов развития, вероятно, казалась задачей студенческого уровня, с которой он мог бы справиться на уроках своего отца в колледже Ломбарда. Когда нужно найти величину некоторой переменной, требуется сначала дать ей обозначение, затем выразить все, что известно об этой переменной и ее связях с другими переменными в виде математических уравнений, и, наконец, если хватит терпения и уравнений, их удастся решить и получить значение нужной переменной.

В примере Райта нужная неизвестная величина (показанная на рис. 15) была обозначена *d* — воздействие внутриутробных факторов развития (*development*) на появление белой окраски. Другие каузальные переменные в уравнении Райта включали *h* — наследственные (*hereditary*) факторы, также неизвестные. Наконец (и в этом проявляется гениальность Райта), он показал, что, если бы нам были известны каузальные переменные на рис. 11, мы могли бы предсказать корреляции в данных (не показанных на диаграмме) на основе простого графического правила.

Сьюалл Райт был первым, кому удалось разработать математический метод для ответа на каузальные вопросы исходя из данных — путевых диаграмм. Сильнее его любви к математике была только его страсть к морским свинкам.

Это правило перебрасывает мост от глубокого, скрытого мира причин во внешний, очевидный мир корреляций. Это была попытка установить связь между причинностью и вероятностью, самое раннее преодоление пространства между первой и второй ступенью Лестницы Причинности. Построив этот

мост, Райт мог двигаться по нему и обратно, от корреляций, вычисляемых на основе данных (ступень первая), к скрытым каузальным переменным d и h (ступень вторая). Он достиг этого, решая алгебраические уравнения. Такая идея, скорее всего, представлялась Райту очень простой, но она оказалась революционной, потому что это было первым доказательством, что мантра «Корреляция не подразумевает причинно-следственных связей» должна уступить место утверждению «Некоторые корреляции как раз подразумевают причинно-следственные связи». В заключение Райт продемонстрировал, что гипотетические факторы внутриутробного развития влияют на окраску сильнее, чем наследственность. В случайно скрещивающейся популяции морских свинок 42% изменчивости окраски обусловлено генетикой, а 58% — факторами внутриутробного развития. По контрасту в высоко инбредной линии только 3% изменчивости в расположении белой окраски по частям тела определялась наследственностью, а 92% — факторами развития. Иными словами, 20 поколений близкородственных скрещиваний почти элиминировали наследственную изменчивость, но факторы, действующие во время развития плода, сохранились.

Как ни интересен этот результат, ключевым моментом для нашей истории является то, каким образом Райт решил данную задачу. Путевая диаграмма на рис. 11 — это дорожная карта, которая объясняет нам, как перемещаться по мосту между первой и второй ступенью. Это целая научная революция в одной картинке — и с умилительными морскими свинками в придачу! Обратите внимание, что путевая диаграмма показывает все мыслимые факторы, способные влиять на окраску детеныша морской свинки. Буквы *D*, E и H относятся к факторам внутриутробного развития, средовым влияниям и наследственным факторам соответственно. Каждый родитель (отец и мать) и каждый потомок (*O* и *O*’), испытывает влияние своего набора факторов *D*, *E* и *H*. У двух потомков общие средовые факторы, но различные истории внутриутробного развития. Диаграмма включает новые в то время для науки идеи менделевской генетики: наследственность

потомка определяется сперматозоидом и яйцеклеткой его родителей (*G* и *G*”), а их наследственный материал, в свою очередь, определяется наследственностью самих родителей (*H*” и *H*””) посредством некоего процесса перемешивания, который на ту пору не был известен (ДНК тогда еще не открыли). Было понятно, впрочем, что перемешивание включает некоторый элемент случайности (обозначенный на диаграмме как «Случайность»).

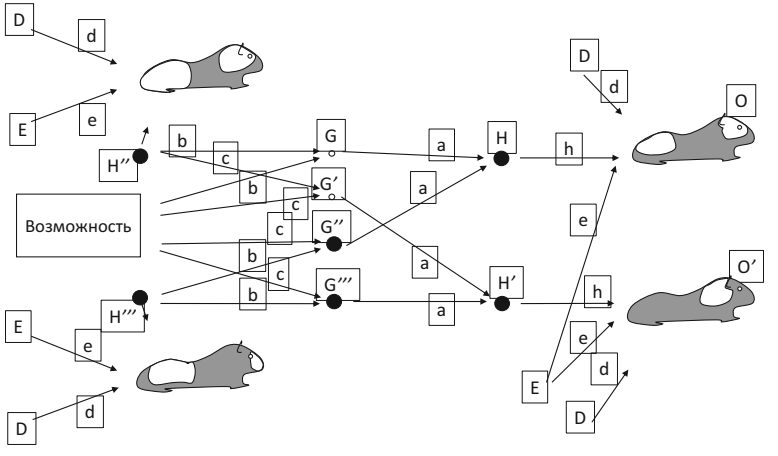


Рис. 11. Первая путевая диаграмма Сьюалла Райта, показывающая все факторы, влияющие на окраску меха у морских свинок: *D* — факторы внутриутробного развития (от зачатия до рождения); *E* — средовые факторы (после рождения); *G* — генетические факторы от каждого из родителей; *H* — объединенные наследственные факторы от обоих родителей, *O*, *O*’ — потомство. Целью анализа была оценка силы воздействия факторов *D*, *E*, *H* (на диаграмме приведенных как *d*, *e*, *h*).

Один момент диаграмма не отражает прямо — разницу между обычной и инбредной семьями. В последней будет сильная корреляция между наследственностью отца и матери, что Райт отметил обоюдонаправленной стрелкой между *H*” и *H*””. Кроме этой, все остальные стрелки на диаграмме направлены в одну сторону и ведут от причины к следствию.

Например, стрелка от *G* до *H* означает, что наследственный материал сперматозоида отца имеет прямое каузальное влияние на наследственность потомства. Отсутствие стрелки от *G* до *H*' означает, что сперматозоид отца, давший жизнь потомку *O*, не влиял каузально на потомка *O*'.

Эти буквы, называемые путевыми коэффициентами, отражают силы каузальных воздействий, которые Райт хотел найти. Грубо говоря, путевой коэффициент отражает долю изменчивости в конечной переменной, которая определяется исходной переменной. Так, достаточно очевидно, что 50% наследственности любого потомка передается от каждого из его двух родителей, поэтому *a* должно быть равно ½ (по техническим причинам Райт предпочитал брать квадратный корень, так чтобы $a = 1/\sqrt{2}$ и $a^2 = ½$). Такая интерпретация путевых коэффициентов, в терминах доли изменчивости, объясняемой данной переменной, в те времена была разумной. Современная причинная интерпретация иная: путевые коэффициенты представляют собой результаты гипотетического воздействия исходной переменной. Однако появления концепции воздействия в 40-х годах XX века нужно было ждать еще долго, и Райт, который написал свою статью в 1920 году, не мог ей воспользоваться. К счастью, в простых моделях, проанализированных им тогда, обе интерпретации приводят к одинаковым результатам.

Я хочу подчеркнуть, что путевая диаграмма не просто красивая картинка, это мощный вычислительный аппарат, потому что правило для подсчета корреляций (мост со второй на первую ступень) включает прослеживание путей, соединяющих две переменные между собой, и перемножение коэффициентов, встреченных по пути. Обратите также внимание, что опущенные на рисунке стрелки на самом деле выражают более важные допущения, чем те, которые на нем присутствуют. Не изображенная стрелка означает, что каузальное воздействие равно нулю, в то время как присутствующая стрелка ничего не говорит нам о силе воздействия (если только мы априорно не придадим путевому коэффициенту определенное значение).

Работа Райта была настоящим прорывом и заслуживает упоминания в качестве эпохального результата в биологии.

Несомненно, это важнейшая веха в истории науки о причинности. Рис. 11 — первая опубликованная каузальная диаграмма, первый шаг XX столетия на вторую ступень Лестницы Причинности, и шаг не робкий, а уверенный и обдуманный! На следующий год Райт опубликовал намного более общую работу под названием «Корреляция и причинность», объясняющую, как путевой анализ работает на другом материале, не только на морских свинках. Не могу представить, какую реакцию на свою публикацию ожидал Райт, но то, что воспоследовало, определенно ошеломило его. Это было опровержение, опубликованное в 1921 году неким Генри Найлзом, учеником американского статистика Раймонда Пирла, который, в свою очередь, был учеником Карла Пирсона, крестного отца статистики.

Академический мир полон цивилизованного людоедства, и мне за свою в основном тихую научную карьеру тоже приходилось испытывать его на собственной шкуре, но все же мне редко попадались настолько злобные критики, как Найлз. Он начинает с длинной серии цитат из своих героев, Карла Пирсона и Фрэнсиса Гальтона, доказывая избыточность или даже бессмысленность термина «причина». Он делает вывод: «Противопоставление „причинности" и „корреляции" необоснованно, потому что причинность — это просто совершенное проявление корреляции». В этом предложении он прямо повторяет то, что Пирсон писал в своей «Грамматике науки».

Далее Найлз старается принизить всю методологию Райта. Он пишет: «Главная ошибка этого метода — предположение, что возможно априори задать относительно простую графическую схему, которая будет верно отражать пути воздействия нескольких переменных друг на друга и на общий результат». Наконец, Найлз разбирает несколько примеров и, путаясь в расчетах, поскольку не дал себе труда разобраться в правилах, установленных Райтом, приходит к противоположным выводам. В итоге он заявляет: «Таким образом, мы заключаем, что с точки зрения философии основания метода путевых коэффициентов ложны, в то время как на практике результаты

применения его там, где возможна проверка, доказывают его совершенную ненадежность».

С научной точки зрения тратить время на детальный разбор опровержения Найлза, вероятно, не стоит, но его статья очень важна для нас, историков науки о причинности. Во-первых, она бесхитростно отражает отношение большинства ученых того поколения к причинности и тотальную власть его наставника Карла Пирсона над научными умами того времени. Во-вторых, возражения Найлза мы продолжаем слышать и сегодня. Конечно, иногда ученые не представляют с точностью всю сложную сеть взаимоотношений между изучаемыми переменными. В этом случае, предполагал Райт, мы можем использовать диаграмму в исследовательском режиме; мы можем постулировать определенные причинно-следственные отношения и рассчитать предсказанные корреляции между переменными. Если они противоречат объективным данным, у нас появляется свидетельство, что отношения, допущенные нами, ложны. Этот способ применения путевых диаграмм, вновь открытый в 1953 году Гербертом Саймоном (ставшим в 1978 году лауреатом Нобелевской премии по экономике), вдохновил множество исследований в общественных науках.

Хотя нам и не нужно знать все причинно-следственные взаимоотношения между интересующими нас переменными и мы в силах делать некоторые выводы, обладая только частичной информацией, Райт подчеркивает один момент с абсолютной четкостью: каузальные выводы невозможно сделать, не имея каузальной гипотезы. Это перекликается с теми выводами, которые мы сделали в главе 1: невозможно ответить на вопрос второй ступени Лестницы Причинности исключительно на основе данных первой ступени. Иногда меня спрашивают: не делает ли это каузальные умозаключения тавтологичными, замкнутыми сами на себя? Разве тем самым вы не предполагаете именно то, что хотите доказать? Правильный ответ — нет. Объединяя очень приблизительные, качественные и очевидные предположения (например, что цвет меха у потомства не влияет на цвет меха родителей) с данными по морским

свинкам за 20 лет наблюдений, Райт получил количественный и совершенно неочевидный результат: окраска меха на 42% определяется наследственностью.

Получить неочевидный результат из очевидных данных — это не тавтология, это научный триумф, заслуживающий, чтобы ему воздали соответствующие почести. Вклад Райта уникален, потому что информация, приведшая к умозаключению (о наследственной компоненте в 42%) была на двух разных и почти несовместимых математических языках: языке диаграмм, с одной стороны, и языке данных — с другой. Еретическая идея объединения качественной «путевой» информации и количественной информации данных (два чуждых друг другу языка!) была чудом, которое привлекло меня, специалиста по компьютерным наукам, к этой проблематике. Многие люди до сих пор повторяют ошибку Найлза, думая, что цель каузального анализа — доказать, что *X* — это причина *Y*, или просто найти причину *Y* с нуля. Это проблема каузальных открытий, которая была моей честолюбивой мечтой еще в те времена, когда я впервые погрузился с головой в графическое моделирование, и до сих пор остается областью активного научного поиска. Напротив, исследования Райта, как главы этой книги, сосредоточены на том, чтобы представить правдоподобные представления о причинно-следственных связях с помощью какого-либо математического языка, объединить их с эмпирическими данными и ответить на вопросы о причинности, имеющие практическое значение. Райт с самого начала понимал, что каузальные открытия, поиск причин — дело намного более сложное, если вообще реальное. В своем ответе Найлзу он пишет: «Автор [т. е. сам Райт] никогда не претендовал на то, что теория путевых коэффициентов может дать нам общую формулу для выяснения причинно-следственных взаимодействий. Он хотел бы подчеркнуть, что сочетание знаний о корреляциях со знанием причинно-следственных связей для получения конкретных результатов не имеет ничего общего с выведением причинно-следственных взаимоотношений из корреляций, о котором пишет Найлз».

E pur si muove (и все-таки она вертится)

Если бы я был профессиональным историком, я бы остановился на этом месте. Но, поскольку я обещался быть историком-вигом, мне не удастся сдержать восхищения точностью слов Райта в цитате, приведенной в конце предыдущего раздела, которые не устарели за 90 лет с тех пор, как он высказал их впервые, и которые в основном и определили парадигму современного каузального анализа.

Мое восхищение точностью формулировки Райта уступает только восхищению его смелостью и целеустремленностью. Только представьте себе ситуацию, сложившуюся в 1921 году. Математик-самоучка в одиночку противостоит гегемонии всего статистического истеблишмента. Они говорят ему: «Ваш метод основан на полном непонимании природы причинности в научном смысле». Он стоит на своем: «Вовсе нет! Мой метод позволяет получать важные результаты и идет в этом дальше, чем все, что смогли придумать вы». Они говорят: «Наши великие гуру уже рассматривали эти вопросы 20 лет назад и решили, что то, что ты делаешь, лишено всякого смысла. Ты просто объединяешь корреляции с корреляциями и получаешь снова корреляции. Когда вырастешь — поймешь». А он продолжает: «Я не пытаюсь опровергнуть ваших гуру, но лопата — это лопата. Мои путевые коэффициенты — это не корреляции. Это нечто совершенно иное — это каузальные воздействия».

Представьте, что вы снова в детском саду и все дети над вами смеются, потому что вы считаете, что $3 + 4 = 7$, в то время как любому ребенку известно, что $3 + 4 = 8$. Вы идете к воспитательнице — а она тоже уверяет вас, что $3 + 4 = 8$. Удалось бы вам не заплакать и не решить, что, наверное, это с вами что-то не то? В таких ситуациях даже самые сильные духом люди начинают сомневаться в истинности своих убеждений. Я сам был в таком детсаду, я знаю.

Но Райт не сдался. И это был не простой арифметический вопрос, в котором возможна независимая верификация. Ранее только философы осмеливались иметь собственное мнение о природе причинности. Откуда у Райта взялась эта внутрен-

няя убежденность, что он на верном пути, а вся остальная группа детсада заблуждается? Может быть, то, что он вырос на Среднем Западе и учился в маленьком, никому не известном колледже, приучило его полагаться на собственные силы и дало понять, что самые надежные знания — это те, которые ты добываешь сам.

Одна из первых прочитанных мною в школе книг о науке рассказывала, как инквизиция заставила Галилея прилюдно отречься от учения о том, что Земля вращается вокруг Солнца, но после отречения тот упрямо прошептал: «И все-таки она вертится!» Вряд ли в мире есть ребенок, который, прочитав эту историю, не был вдохновлен смелостью и верностью Галилея своим убеждениям. Однако, как бы мы ни восхищались его позицией, сложно не думать о том, что он мог опираться по крайней мере на свои точные астрономические наблюдения. У Райта под рукой были только непроверенные выводы, например, что факторы внутриутробного развития отвечают за 58%, а не за 3% изменчивости окраски. Не имея ничего, на что можно было бы опереться, кроме внутреннего убеждения, что путевые коэффициенты способны рассказать нам то, чего не знают корреляции, он тем не менее объявил: «И все-таки она вертится!»

Коллеги говорят мне, что, когда истеблишмент в области искусственного интеллекта боролся с байесовскими сетями (см. главу 3), я действовал упрямо, самоуверенно и бескомпромиссно. В самом деле, я помню, что был совершенно уверен в верности своего подхода и не колебался ни на йоту. Но на моей стороне была теория вероятностей. Райт же не мог опереться даже на подходящую теорему. Ученые его времени отказались от причинности, поэтому никакого теоретического фундамента под свою работу он подвести не мог. Не мог он и опереться на авторитетные мнения, как тот же Найлз, потому что цитировать было некого: великие гуру вынесли свои окончательные вердикты еще десятилетиями ранее.

Однако у Райта было и утешение, был знак, что он на верном пути — понимание, что его метод дает ответы на вопросы, на которые нельзя ответить никак иначе. Одним из таких

вопросов было определение относительной силы влияния нескольких факторов. Другой замечательный пример — в его статье «Корреляция и причинность» за 1921 год, где выясняется, как дополнительный день в утробе матери повлияет на вес новорожденной морской свинки. Ниже я разберу ответ Райта детально, чтобы показать красоту его метода и порадовать тех читателей, которые хотели бы видеть, как работает путевой анализ с математической стороны.

Обратим внимание, что мы не ответим на этот вопрос прямо, потому что не в силах взвесить морскую свинку еще в утробе. Мы, однако, способны сравнить вес при рождении у морских свинок, беременность матери которых длилась, скажем, 66 дней, с теми, которые провели в утробе 67 дней. Райт отметил, что, если беременность длилась на один день дольше, новорожденные свинки в среднем весили больше на 5,66 грамма. Можно наивно предположить, что за последний день в животе матери каждый детеныш морской свинки поправляется на эти 5,66 грамма.

«Неверно!» — говорит Райт. Детеныши обычно появляются на свет позже не просто так, а по определенной причине: в таких пометах обычно меньше детенышей. Это значит, что в течении всей беременности условия развития у них были лучше. Новорожденная морская свинка из помета, в котором было только трое детенышей, уже на 66-й день весит больше, чем из помета, в котором их было пятеро. Таким образом, разница в весе при рождении объясняется двумя разными причинами и нам надо их распутать. Сколько из дополнительных 5,66 грамма детеныш набирает за счет того, что проводит в матке на день больше, а сколько — за счет того, что у него меньше конкурентов?

Райт ответил на этот вопрос, начертив путевую диаграмму (рис. 12).

X — это вес детеныша при рождении. P и Q — два фактора, о которых известно, что они влияют на вес детеныша: P — продолжительность беременности, а Q — скорость роста в утробе матери. L — это число детенышей в помете, которое влияет сразу и на P, и на Q (при большом помете детеныши растут

медленнее, а беременность длится меньше). Важно обратить внимание, что *X*, *P* и *L* можно измерить для каждого животного в отдельности, а *Q* — нельзя. Наконец, *A* и *C* — внешние причины, по которым у нас нет данных (т.е. наследственные и средовые факторы, влияющие на продолжительность беременности и скорость внутриутробного развития вне зависимости от числа детенышей в помете). Важное предположение, что эти факторы не зависят друг от друга, выражается отсутствием стрелки между ними, равно как и причины, влияющей на оба этих фактора.

Теперь можно сформулировать вопрос, стоявший перед Райтом: каково прямое влияние продолжительности беременности *P* на вес при рождении *X*? Данные (5,66 грамма за день) ничего не говорят нам о прямом влиянии — они дают нам только корреляцию, смещенную за счет влияния числа детенышей в помете *L*. Чтобы найти прямое влияние, мы должны устранить это смещение.

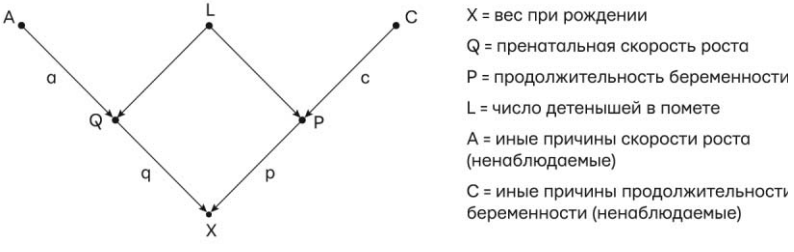


Рис. 12. Диаграмма причинности (путевая) для примера с весом при рождении

На рис. 12 прямое влияние обозначено путевым коэффициентом p, соответствующим пути $P \rightarrow X$. Смещение за счет числа детенышей в помете соответствует пути $P \leftarrow L \rightarrow Q \rightarrow X$. А теперь в игру вступает магия алгебры: величина смещения равна произведению путевых коэффициентов вдоль по данному пути (иными словами, l умножить на l’ и умножить на *q*). Общая корреляция тогда равна просто сумме путевых коэффициентов по обоим путям: алгебраически $p + (\mathrm{l} \cdot \mathrm{l}' \cdot q) = 5{,}66$ грамма

в день. Если бы мы знали величину путевых коэффициентов q, l и l', мы бы могли рассчитать второе слагаемое и вычесть его из 5,66, получив p. Но мы их не знаем, потому что Q, например, невозможно измерить. Но именно здесь и проявляется гениальность метода путевых коэффициентов. Метод Райта расписывает, как выразить каждую из посчитанных корреляций в соответствующих терминах. Сделав это для каждой из измеренных пар (P, X), (L, X) и (L, P), мы получаем три уравнения, которые решаются алгебраически для неизвестных путевых коэффициентов, p, l' и $(l \cdot q)$. После этого задача решена, желаемая величина p найдена.

Сегодня мы можем обойтись вообще без математики и рассчитываем p посредством беглого изучения диаграммы. Но в 1920 году это был первый случай, когда математику призвали объединить корреляции и причинность. И это сработало! Райт вычислил, что p равно 3,34 грамма в день. Другими словами, если все другие переменные (A, C, L, Q) остаются постоянными и только срок беременности увеличится на один день, средний рост веса при рождении составит 3,34 грамма. Заметим, что этот результат имеет внятный биологический смысл. Он говорит нам, с какой скоростью детеныши растут в каждый день внутриутробного развития. Число 5,66, напротив, биологически бессмысленно, потому что оно смешивает два разных процесса, один из которых не каузальный, а антикаузальный (или диагностический): это связь $P \leftarrow L$.

Приведенный пример преподает нам два урока. Первый: причинный анализ позволяет нам находить численные выражения реальных процессов в реальном мире, а не только структуры данных. Детеныши растут со скоростью 3,34 грамма в день, а не 5,66 грамма в день. Урок второй: следили вы за математикой или нет, но в путевом анализе мы делаем выводы об индивидуальных причинно-следственных отношениях, изучая диаграмму в целом. Чтобы оценить каждый индивидуальный параметр, может понадобиться структура всей диаграммы.

В воображаемом мире, где наука развивается логично, ответ Райта Найлзу должен был бы вызвать всеобщий научный восторг, а затем его методы с энтузиазмом стали бы применять

другие ученые и статистики. Но судьба распорядилась иначе. «Одна из загадок истории науки в период с 1920 по 1960 годы — это практически полное отсутствие применения путевого анализа, за исключением самого Райта и селекционеров животных, — писал один из коллег Райта генетик Джеймс Кроу. — Хотя Райт продемонстрировал много примеров возможного применения своего метода, ни по одному из предложенных им путей никто не пошел».

Кроу не знал об этом, но такое загадочное умолчание коснулось и общественных наук. В 1972 году экономист Артур Гольдбергер оплакивал «постыдную неизвестность» работ Райта в тот период и отмечал, с энтузиазмом новообращенного, что «подход [Райта] стал искрой, воспламенившей нынешний интерес к каузальным моделям в социологии. Ах, если бы мы могли обратиться к современникам Райта и спросить — почему вы не обратили внимания? Кроу дает такой ответ: „путевой анализ не годится для программ-„консервов". Пользователь должен самостоятельно сформировать гипотезу и создать годную диаграмму из множества причинных последовательностей"». Действительно, Кроу указал на важный момент: путевой анализ, как и любое упражнение в области причинно-следственных связей требует умения научно мыслить. Статистика же, как это часто случается, не поощряет его, способствуя появлению программ-«консервов», применяемых механически. Ученые всегда будут предпочитать рутинные вычисления на основе данных методам, которые бросают вызов их научным познаниям.

Рональд Эйлмер Фишер, непререкаемый авторитет в области статистики в поколении после Гальтона и Пирсона, характеризует эту разницу лаконично. В 1925 году он пишет: «Статистику можно назвать… наукой о методах редукции данных». Обратите внимание на слова «методы», «редукция» и «данные». Райту претило представление о статистике как только о собрании методов — Фишеру оно было по душе. Причинно-следственный анализ, подчеркнем, не сводится к данным: в ход анализа мы должны инкорпорировать некоторые представления о процессах, которые приводят к появлению этих данных, и тогда

мы получаем в результате нечто, что исходно в наших данных не содержалось.

Но в одном Фишер был прав: если убрать из статистики причинность, редукция данных — это все, что вам остается.

Хотя Кроу и не упоминает этого, биограф Райта Уильям Провин указывает еще на один фактор, который мог повлиять на недостаток поддержки путевого анализа. С середины 30-х годов XX века Фишер считал Райта своим врагом. Я ранее цитировал воспоминания Юла о том, как отношения с Пирсоном резко становились натянутыми, если кто-то не соглашался с ним, и невозможными — если Пирсона критиковали. Совершенно то же самое справедливо и в отношении Фишера. Последний устраивал продолжительные вендетты всем, с кем был не согласен, включая Пирсона, его сына Эгона, Ежи Неймана (о них обоих будет подробнее в главе 8) и, конечно, Райта.

Главной точкой соперничества Фишера и Райта был не путевой анализ, а эволюционная биология. Фишер был не согласен с теорией Райта (называемой генетическим дрейфом), согласно которой вид может эволюционировать очень быстро, когда проходит через популяционное бутылочное горлышко. Детали этого спора выходят за рамки данной книги, и заинтересованный читатель должен обратиться к работе Провина. Важно здесь, однако, следующее: с 20-х до 60-х годов XX века научный мир был преимущественно повернут лицом к Фишеру, как к оракулу статистических знаний. И уверяю вас, что он никогда не сказал ни одного доброго слова про путевой анализ.

В 1960-х все стало понемногу меняться. Группа представителей общественных наук, включающая Отиса Дункана, Хьюберта Блалока и экономиста Артура Гольдбергера (упомянутого ранее), заново открыла путевой анализ как метод предсказания результатов социальной и образовательной политики. По иронии судьбы Райта приглашали выступить перед влиятельной группой эконометриков, комиссией Каулза, в 1947 году, но ему совершенно не удалось разъяснить им, в чем смысл путевых диаграмм. Только когда экономисты сами додумались до подобных идей, на короткое время удалось установить контакт. Судьба путевого анализа в экономике

и в социологии двигалась по разным траекториям, но каждая из них вела к предательству идей Райта. Социологи переименовали путевой анализ в уравнения структурного моделирования (*Structural Equation Modeling; SEM*), полюбили диаграммы и активно их использовали до 70-х годов прошлого века, когда компьютерная программа LISREL автоматизировала подсчет путевых коэффициентов (в некоторых случаях). Райт мог бы предсказать то, к чему это привело: путевой анализ превратился в рутинный метод, а исследователи стали пользователями программы, которых слабо интересует, что у нее внутри. В конце 1980-х обращенный к научному обществу призыв (статистика Дэвида Фридмана) объяснить допущения, стоящие за уравнениями структурного моделирования, был проигнорирован, а некоторые эксперты по SEM даже отрицали, что этот метод как-то связан с причинностью.

В экономике алгебраическая часть путевого анализа получила известность как системы одновременных уравнений (без общепринятого сокращения). Экономисты почти никогда не использовали путевые диаграммы и продолжают обходиться без них и по сей день, полагаясь взамен на числовые уравнения и матричную алгебру. Прямое последствие этого подхода в том, что, поскольку алгебраические равенства не направлены (т.е. $x = y$ — это то же самое, что и $y = x$), у экономистов не было обозначений для различения причинных и регрессионных уравнений, поэтому на вопросы, связанные с выбором стратегии, не получалось ответить даже после того, как уравнения были решены. Вплоть до 1995 года большинство экономистов избегали прямо придавать своим уравнениям каузальный или контрфактивный смысл. Даже те из них, кто использовал структурные уравнения для обоснования выбора стратегических решений, неизлечимо боялись диаграмм, хотя те могли бы избавить их от многих страниц лишних вычислений. Неудивительно, что многие экономисты и сегодня стоят на точке зрения «в данных уже содержится все».

По всем этим причинам многообещающие перспективы путевых диаграмм оставались реализованы только отчасти, по крайней мере до 90-х годов прошлого века. В 1983 году

Райта снова лично вызвали на ринг защищать их, на этот раз в «Американском журнале генетики человека». В то время, когда он писал эту статью, ему уже было за 90. Читать его эссе 1983 года одновременно и радостно, и больно — оно о том же, о чем он писал в 1923 году. Много ли в истории науки случаев, когда нам выпадала честь вновь выслушать создателя великой теории через 60 лет после того, как он впервые изложил ее на бумаге? Это примерно как если бы Чарлз Дарвин восстал из могилы, чтобы поприсутствовать на «обезьяньем процессе» Скоупса в 1925 году. Но это и трагично, потому что за эти 60 лет теория должна была расти, развиваться, цвести, а вместо этого она осталась практически на том же уровне, что и в 1920-х.

Написать статью Райта побудила критика путевого анализа, которую ранее опубликовали в том же журнале Самуэль Карлин (математик Стэнфордского университета, награжденный в 1989 году Национальной научной медалью, внесший фундаментальный вклад в экономику и популяционную генетику) и два его соавтора. Нас здесь интересуют два аргумента Карлина.

Во-первых, путевой анализ вызывает у Карлина возражения по причине, которую Найлз не упомянул: он предполагает, что все взаимоотношения между любыми переменными в путевой диаграмме линейны. Это допущение позволило Райту описать каузальные взаимодействия с помощью только одного числа — путевого коэффициента. Если бы уравнения были нелинейны, то тогда воздействие на Y от изменения X на единицу зависело бы от текущего значения X. Ни Карлин, ни Райт не знали, что до появления общей теории нелинейности осталось совсем немного (ее всего три года спустя разработает звезда моей лаборатории Томас Верма).

Но самое интересное возражение Карлина — как раз то, которое он считал важнейшим: «Наконец, с наибольшим успехом исследователь может выбрать подход, не предполагающий модели вообще, приводящий к пониманию данных интерактивно, используя ряд отображений, индексов и контрастов. Этот подход подчеркивает концепцию надежности данных в интерпретации

результатов». В одном этом коротком отрывке Карлин выражает, сколь мало изменилось со времен Пирсона и насколько влиятельной его идеология оставалась даже в 1983 году. Он говорит, что в самих данных уже заключена вся научная мудрость; их нужно только уметь умаслить и сделать им массаж (с помощью отображений, индексов и противопоставлений), и они сами выронят жемчуг мудрости вам в руки. Нашим аналитикам нет нужды принимать во внимание процессы, которые привели к появлению этих данных. У нас все получится ровно так же, и даже лучше с подходом, «не предполагающим никакой модели вообще». Если бы Пирсон жил сегодня, в эпоху больших данных, он сказал бы ровно это: все ответы уже содержатся в самих данных. Конечно, утверждения Карлина нарушают все, о чем мы говорили в первой главе. Чтобы говорить о причинности, нам требуется ментальная модель реального мира. «Безмодельный подход» может привести нас на первую ступень Лестницы Причинности, но никак не дальше.

Райт, надо отдать ему должное, прекрасно понимал, как велики ставки, и написал недвусмысленно: «Заявляя, что подход, не предполагающий модели вообще — наилучшая альтернатива… Карлин с соавторами хотят не просто изменить метод путевого анализа, но лишить его цели и оценки относительной важности различных причин. Этот анализ невозможен без модели. Они предлагают тем, кому хочется провести такую оценку, подавить свое желание и заняться чем-нибудь другим».

Райт понимал, что защищает саму суть научного подхода и интерпретации данных. Сегодня я бы дал энтузиастам больших данных, избегающим моделей, тот же совет. Конечно, замечательно попытаться выудить всю информацию, которую данные способны нам сообщить, но надо понимать, насколько далеко это позволит нам уйти. А уйти оно позволит не дальше первой ступени Лестницы Причинности и никогда не сможет дать ответ даже на такой простой вопрос: какова относительная важность различных действующих факторов?

E pur si muove!

От объективности к субъективности: мост, переброшенный Байесом

Еще одна тема из отповеди Райта может намекнуть на другое обстоятельство, по которой статистики сопротивлялись причинности. Он многократно утверждает, что не хотел бы, чтобы путевой анализ стал «стереотипным». Буквально по Райту: «Нестереотипный подход путевого анализа принципиально отличается от стереотипных моделей описания, созданных для того, чтобы избегать малейших отклонений от полной объективности».

Что он имеет в виду? Во-первых, то, что путевой анализ должен быть основан на личном понимании причинно-следственных процессов, отраженных в каузальных диаграммах. Он не может быть редуцирован до механических процедур вроде тех, что описываются в справочниках по статистике. Для Райта рисование путевой диаграммы — не упражнение в статистике; это упражнение в генетике, экономике, психологии или любой другой области, экспертом в которой является ученый.

Во-вторых, Райт прослеживает связь очарования «безмодельных» методов с их объективностью. Для статистики объективность действительно была святым Граалем с самого первого дня, или же с 15 марта 1834 года, когда было основано Лондонское статистическое общество. В его уставе сказано, что данные во всех случаях имеют приоритет над мнениями и интерпретациями. Данные объективны — мнения субъективны. Эта парадигма возникла задолго до Пирсона. Борьба за объективность — принцип выведения умозаключений только на основе данных и экспериментов — была важнейшим моментом в том, как наука определяла сама себя со времен Галилея.

В отличие от корреляций и большинства других инструментов общепринятой статистики, каузальный анализ требует от пользователя субъективной заинтересованности. Ему потребуется нарисовать каузальную диаграмму, отражающую его качественные представления, или, скорее, консенсусные представления исследователей в его области науки, о топо-

логии происходящих в данном случае каузальных процессов. Он должен забыть о многовековой догме объективности для ее же пользы. Там, где дело касается причинности, одно зерно разумной субъективности говорит нам больше о реальном мире, нежели любые объемы объективности.

Абзацем выше я сказал, что «большинство» инструментов статистики стремится к полной объективности. Для этого правила, однако, есть одно серьезное исключение. Область статистики, именуемая байесовой статистикой, за последние примерно 50 лет достигла значительной популярности. Когда-то ее едва ли не проклинали, но теперь это нечто совершенно общепринятое, и на конференции по статистике за все время работы уже не услышать ни одного спора между «байесианцами» и «частотниками», хотя в 1960-х и 1970-х они гремели.

Прототип байесовского анализа таков: предварительные представления + новые данные = пересмотренные представления. Представьте, что вы подбросили монету десять раз, и девять из них она выпадала орлом. Ваша уверенность в том, что монета не фальшивая, поколеблена, но насколько? Традиционный статистик скажет: «При отсутствии дополнительных данных я предположил бы, что эта монета с грузом, и я поставлю девять против одного, что в следующий подброс она выпадет орлом». Байесовский статистик возразит: «Подождите. Мы должны учесть уже имеющиеся данные о происхождении монеты». Откуда она взялась: из сдачи в гастрономе или из кармана мошенника? Если это обычный гривенник, то выпадение девяти орлов подряд не должно вызывать у нас настолько сильных подозрений. И наоборот, если мы уже подозревали, что с монетой что-то не так, мы заключим с большей уверенностью, что девять орлов — это серьезное нарушение случайного распределения.

Байесова статистика дает нам объективный способ объединить результаты наблюдений с нашими предварительными знаниями (или субъективными представлениями), чтобы получить пересмотренные представления и, следовательно, пересмотренные предсказания о том, как поведет себя монета при

следующем подбрасывании. Однако чего частотники не могли простить, так это того, что байесианцы позволили мнению, в виде субъективной вероятности, проникнуть в стерильное царство статистики. Признания большинства удалось заслужить только очень постепенно, когда байесовский анализ проявил себя как превосходный инструмент для решения множества задач, таких разных, как предсказание погоды и отслеживание вражеских подводных лодок. Вдобавок во множестве случаев можно доказать, что влияние предварительных представлений тает с ростом массива данных, так что в конце остается чисто объективный вывод.

К сожалению, то, что общепринятая статистика смирилась с байесовской субъективностью, никак не повлияло на ее отношение к субъективности каузальной, требующейся для составления путевых диаграмм. Почему? Ответ лежит в плоскости великого языкового барьера. Чтобы озвучить субъективные предположения, байесова статистика все-таки использует язык вероятностей — родной язык Гальтона и Пирсона. Предположения о причинно-следственных связях, однако, требуют более богатого языка (например, диаграммы), который одинаково чужд и байесианцам, и частотникам. Взаимопонимание, к которому пришли последние, показывает, что преодолеть философские барьеры помогает добрая воля и общий язык. Языковые барьеры не так легко перепрыгнуть.

Более того, субъективная компонента в каузальной информации далеко не всегда уменьшается со временем, даже с ростом объема данных. Два человека, у которых два разных взгляда на причинность одного явления, могут анализировать один и тот же набор данных и никогда не прийти к общему результату, как бы велик объем данных ни был. А это пугающая перспектива для адвокатов научной объективности, и она объясняет их отказ принять неизбежность необходимости полагаться на субъективную каузальную информацию.

Положительная сторона этой проблемы в том, что причинностное умозаключение объективно в одном принципиально важном смысле: если двое ученых согласны в своих предполо-

жениях, оно позволяет со 100%-ной объективностью интерпретировать новую входящую информацию (данные). В этом оно совпадает с байесовским умозаключением.

Так что пытливый читатель, вероятно, не удивится тому, что я пришел к теории причинности окольным путем, который начинался с байесовских вероятностей и затем шел через байесовские сети. Эту историю я расскажу в следующей главе.

Глава 3
От доказательств к причинам. Преподобный Байес знакомится с мистером Холмсом

Пойдут ли двое вместе,
не сговорившись между собою?
Ревет ли лев в лесу, когда нет перед ним добычи?
Книга пророка Амоса. 3:3

«Элементарно, Ватсон!» — так говорил Шерлок Холмс (по крайней мере, в кино), прежде чем изумить верного помощника характерным и подчеркнуто неэлементарным дедуктивным рассуждением. Но на деле Холмс занимался не просто дедукцией, которая ведет от гипотезы к заключению. Он прекрасно владел искусством индукции, которая работает в противоположном направлении — от улики к гипотезе.

Еще одна известная цитата описывает его образ действий: «Если исключить невозможное, то, что останется, и будет правдой, сколь бы невероятным оно ни казалось». Получив несколько гипотез методом *индукции*, Холмс затем отметал одну за другой, чтобы с помощью *дедукции* (исключения) найти верную. Хотя индукция и дедукция идут рука об руку, первая гораздо загадочнее. Этот факт и позволяет детективам вроде Шерлока Холмса оставаться в деле.

Однако в последние годы эксперты по искусственному интеллекту добились большого прогресса в автоматизации процесса умозаключений, ведущего от улик к гипотезам и подобным же

образом — от следствий к причинам. Мне повезло участвовать в этом процессе на самых ранних стадиях: я разработал один из его базовых инструментов под названием «байесовские сети». Эта глава объясняет, что они собой представляют, рассматривает способы их применения сегодня и обсуждает окольные пути, по которым они привели меня к исследованию причинно-следственных связей.

Bonaparte — компьютер-детектив

17 июля 2014 года рейс MH17 авиакомпании «Малайзия эйрлайнс» вылетел из амстердамского аэропорта Схипхол в Куала-Лумпур. Увы, самолет не добрался до пункта назначения. Через три часа, когда самолет пролетал над Восточной Украиной, его сбили ракетой «земля — воздух» российского производства. Все 298 человек на борту, 283 пассажира и 15 членов экипажа, погибли в авиакатастрофе.

23 июля, когда в Нидерланды были доставлены первые погибшие, объявили днем государственного траура. Но для криминалистов из Нидерландского института судебной экспертизы в Гааге 23 июля стало точкой отсчета. В их задачи входило как можно скорее идентифицировать останки и доставить их близким для похорон. Время поджимало, потому что каждый день неизвестности приносил обездоленным семьям новую боль.

Криминалисты столкнулись со множеством препятствий. Тела были сильно обожжены, и многие хранились в формальдегиде, который разрушает ДНК. Кроме того, поскольку Восточная Украина оставалась территорией военных действий, место авиакатастрофы было доступно не всегда. Останки находили в течение десяти месяцев. К тому же криминалисты не располагали информацией о ДНК жертв по той простой причине, что погибшие не были преступниками. Поэтому приходилось полагаться на частичные совпадения с ДНК родственников.

К счастью, у голландских специалистов был мощный инструмент — новейшая программа под названием *B*onaparte, предназначенная для идентификации жертв катастроф. Эта программа, которую разработали в середине 2000-х ученые

из Университета Неймегена имени святого Радбода Утрехтского, использует байесовские сети, чтобы скомбинировать информацию о ДНК, взятую у нескольких членов семьи.

Отчасти благодаря скорости и точности *B*onaparte голландские криминалисты смогли опознать останки 294 из 298 жертв к декабрю 2014 года. К 2016 году только две жертвы катастрофы (оба граждане Нидерландов) оставались пропавшими без вести.

Байесовские сети, инструмент для машинного рассуждения, лежащий в основе программы *B*onaparte, влияет на нашу жизнь разными способами, о которых большинство людей не имеет представления. Они используются в программах распознавания речи, фильтрах для спама, прогнозах погоды, при оценке потенциальных нефтяных скважин и одобрении медицинских приборов в Управлении по санитарному надзору за пищевыми продуктами и медикаментами. Если вы играете в видеоигры на приставке *X*b*o*X компании «Майкрософт», значит, байесовские сети оценивают ваш уровень. Если у вас есть мобильный телефон, то алгоритмы, которые используются, чтобы выбрать ваш исходящий вызов из тысяч других, кодируются с помощью алгоритма распространения доверия, разработанного для байесовских сетей. Винт Серф, главный пророк Интернета из еще одной небезызвестной компании — Google, — говорит об этом так: «Мы потребляем байесовские методы в огромных объемах».

В этой главе я расскажу историю байесовских сетей с их появления в *X*VIII веке до развития в 1980-х годах, а еще приведу больше примеров того, как они используются сегодня. Они связаны с диаграммами причинности очень простым способом: такая диаграмма — это байесовская сеть, в которой каждая стрелка обозначает прямое причинно-следственное отношение или, по крайней мере, его возможность в направлении этой стрелки. Не все байесовские сети имеют причинно-следственную природу — во многих случаях это не имеет значения. Однако, если вы когда-нибудь захотите задать вопрос второго или третьего уровня на Лестнице Причинности, вам необходимо будет нарисовать диаграмму, обратив самое пристальное внимание на причинно-следственные связи.

Преподобный Байес и проблема обратной вероятности

Томас Байес, в честь которого я назвал сети в 1985 году, даже и не мечтал, что формула, которую он вывел в 1750-х годах, однажды будет использоваться, чтобы идентифицировать жертв катастрофы. Его волновала исключительно вероятность двух событий, одно из которых (гипотеза) происходит после второго (подтвержденного факта). Тем не менее причинность весьма его волновала. Более того, стремление установить причинно-следственные связи было движущей силой для его анализа «обратной вероятности».

Преподобный Томас Байес, пресвитерианский священник, живший с 1702 по 1761 годы, очевидно, был сильно увлечен математикой. Отколовшись от англиканской церкви, он не мог учиться в Оксфорде или Кембридже и вместо этого получил образование в Эдинбургском университете, где, вероятно, немало занимался любимой наукой. После того как Байес вернулся в Англию, он продолжал баловаться математикой и организовывать математические обсуждения.

В статье, опубликованной после его смерти, Байес разобрал задачу, которая была для него идеальной: столкнул математику и теологию. Это произошло в следующих обстоятельствах: в 1748 году шотландский философ Дэвид Юм написал эссе под названием «О чудесах», в котором утверждал, что личное свидетельство никогда не может служить подтверждением для чуда. Чудом, которое Юм имел в виду, было, конечно, воскресение Христа, хотя он был достаточно умен, чтобы этого не сказать (20 годами ранее теолог Томас Вулстон был обвинен в богохульстве и сел в тюрьму за такие утверждения). Главная мысль Юма состояла в том, что наблюдения, которые по природе своей могут быть ошибочными, не способны опровергнуть положение, основанное на законах природы, например: «Мертвые люди остаются мертвыми».

В глазах Байеса это утверждение приводило к естественном вопросу, прямо в духе Холмса: сколько доказательств необ-

ходимо, чтобы убедить нас в том, что события, которые мы считаем невероятными, все же произошли? Если исключить невозможное, то, что останется, и будет правдой, сколько бы невероятным это ни казалось. Когда гипотеза переходит границу между невозможным и невероятным или даже между вероятностью и подлинной уверенностью? Хотя этот вопрос был выражен на языке вероятности, за ним стояли намеренно богословские выкладки. Ричард Прайс, коллега-священник, который нашел эссе в вещах Байеса после его смерти и отправил его в печать с хвалебным вступлением, написанным самолично, выразил эту мысль предельно ясно: «Цель, которую я имею в виду, состоит в том, чтобы показать, по какой причине мы верим, что в порядке вещей существуют неизменные законы, в соответствии с которыми все происходит, и что, таким образом, мироустройство должно быть результатом мудрости и мощи разумной причины, а значит, подтвердить аргумент, основанный на конечных причинах, в пользу существования Всевышнего. Будет легко увидеть, что обратную проблему, решенную в этом эссе, легче применить для этой цели; она показывает нам ясно и точно, каковы основания полагать, что в любом случае каждого конкретного порядка и повторяемости событий этот порядок или повторяемость объясняются стабильной причиной и законами природы, а не случайностями, не подчиненными порядку».

Сам Байес не касался ничего этого в своем тексте; Прайс подчеркнул эти теологические выводы — возможно, чтобы эффект от работы друга был более масштабным. Но оказалось, что Байес не нуждался в помощи. О его работе помнят и ее обсуждают 250 лет спустя, и не из-за теологического значения, а потому, что она показывает: вероятность причины реально вывести из следствия. Если мы знаем причину, легко оценить вероятность следствия — прямую вероятность. Пойти в другом направлении — эту задачу во времена Байеса называли обратной вероятностью — сложнее. Байес не объяснил, почему она

сложнее, — он счел это самоочевидным, доказал возможность ее решить и показал нам, как это сделать.

Чтобы оценить суть этой проблемы, давайте рассмотрим пример, который он сам предложил в работе 1763 года, напечатанной посмертно. Представим, что мы делаем удар кием по бильярдному мячу на столе и стараемся, чтобы он отскочил много раз — так, чтобы у нас не было представления о том, где он окажется. Какова вероятность того, что он остановится через *X* футов от левого края стола? Если мы знаем длину стола и если он абсолютно гладкий и плоский, это очень легкий вопрос (рис. 13а). Так, на 12-футовом столе для снукера вероятность того, что мяч остановится в футе от края, составит $1/12$. На восьмифутовом бильярдном столе вероятность будет $1/8$.

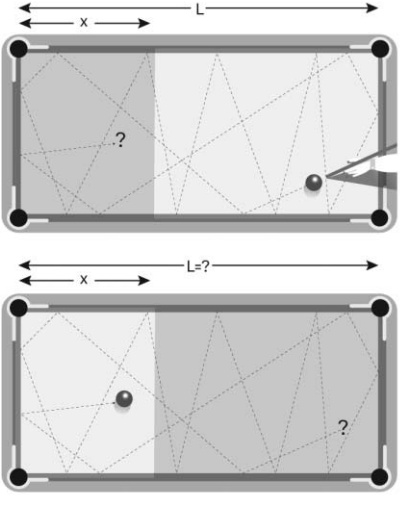


Рис. 13. Пример Томаса Байеса с бильярдным столом: *а* — в первом варианте, с вопросом о прямой вероятности, мы знаем длину стола и хотим вычислить вероятность того, что шар остановится в *x* футах от края; *б* — во втором варианте, с вопросом об обратной вероятности, мы наблюдаем, что шар остановился в *x* футах от конца и хотим оценить вероятность того, что длина стола составляет *L* (источник: рисунок Маян Харел)

Интуитивное понимание физики говорит нам, что в общем, если длина стола составляет L футов, вероятность того, что шар остановится в X футах от края составляет x/L. Чем больше длина стола L, тем ниже вероятность, потому что за право зваться конечным положением шара соревнуются больше позиций. Сдругой стороны, чем больше x, тем выше вероятность, поскольку она включает большее число конечных позиций.

Теперь рассмотрим проблему обратной вероятности. Мы наблюдаем конечное положение шара, в котором $x = 1$ фут от края, но не знаем длины (рис 13б). Преподобный Байес спросил: какова вероятность того, что длина была, скажем, 100 футов? Здравый смысл подсказывает, что длина, вероятнее, составила 50 футов, а не 100, ведь чем длиннее стол, тем труднее объяснить, почему шар оказался так близко к краю. Но насколько это вероятнее? «Интуиция» или «здравый смысл» не дает нам четких указаний.

Почему прямую вероятность (x при известном L) настолько легче оценить в уме, чем вероятность L при известном x? В этом примере асимметрия объясняется тем фактом, что L выступает в роли причины, а x — следствия. Если мы наблюдаем причину, скажем Бобби бросает мяч в окно, большинство может предсказать эффект (мяч, вероятно, разобьет окно). Человеческое познание работает в этом направлении. Но при известном следствии (окно разбито) нам требуется гораздо больше информации, чтобы вывести причину (кто из мальчиков бросил мяч, разбивший окно, или было ли окно вообще разбито мячом). Чтобы учесть все возможные причины, необходим ум Шерлока Холмса. Байес решил удалить эту когнитивную асимметрию и объяснить, как даже обычные люди могут оценить обратную вероятность.

Чтобы посмотреть, как работает метод Байеса, давайте начнем с простого примера о посетителях чайной, о которых у нас есть данные: мы знаем об их предпочтениях. Данные, как нам известно из главы 1, совершенно не в курсе, что существует асимметрия причины и следствия, а значит, с их помощью мы можем найти способ, как разрешить загадку обратной вероятности.

Предположим, что две трети покупателей приходят заказать чай и что половина пьющих чай также заказывают пирожные. Какова будет доля клиентов, которые закажут и чай, и пирожные? В этом вопросе нет подводных камней, и я надеюсь, что ответ почти очевиден.

Поскольку половина двух третей — одна третья, выходит, что одна третья клиентов заказывает чай и пирожные. Чтобы проиллюстрировать это числами, предположим, что мы занесли в таблицу заказы следующих 12 посетителей, которые войдут в дверь.

Как показывает табл. 1, $\frac{2}{3}$ (1, 5, 6, 7, 8, 9, 10, 12) заказали чай и половина из них заказала пирожные (1, 5, 8, 12). Таким образом, доля клиентов, которые заказали и чай, и пирожные действительно равна $\frac{1}{2} \cdot \frac{2}{3} = \frac{1}{3}$, ровно как мы и предсказывали до того, как увидели конкретные данные.

Таблица 1. Вымышленные данные для примера с чаем и пирожными

Посетитель	Чай	Пирожные	Посетитель	Чай	Пирожные
1	Да	Да	7	Да	Нет
2	Нет	Да	8	Да	Да
3	Нет	Нет	9	Да	Нет
4	Нет	Нет	10	Да	Нет
5	Да	Да	11	Нет	Нет
6	Да	Нет	12	Да	Да

Отправная точка для байесовского правила — заметить, что данные можно было проанализировать в обратном порядке, т.е. мы могли бы заметить, что $\frac{5}{12}$ клиентов (1, 2, 5, 8, 12) за-

казали пирожные, а $\frac{4}{5}$ из них (1, 5, 8, 12) заказали чай. Таким образом, доля клиентов, которые заказали и чай, и пирожные, будет вычисляться так: $\frac{4}{5} \cdot \frac{5}{12} = \frac{1}{3}$. Конечно, не случайно у нас получился один и тот же результат; мы просто вычислили одно и то же разными способами. Порядок, в котором клиенты объявляют свои заказы, не играет никакой роли.

Чтобы сделать из этого общее правило, пусть P (T) обозначает вероятность того, что посетитель закажет чай, а P (S) — вероятность того, что он закажет пирожные (помните, что вертикальная линия обозначает «при том что»). Подобным образом, P ($T \mid S$) обозначает вероятность заказа посетителем чая при том, что мы уже знаем о заказе им пирожных.

Сначала мы вычисляем следующее:

$$P \text{ (S and } T) = P \text{ } (S \mid T) \text{ } P \text{ } (T).$$

Второй расчет выглядит так:

$$P \text{ } (S \text{ and } T) = P \text{ } (T \mid S) \text{ } P \text{ } (S).$$

Как говорил Евклид 2 300 лет назад, две величины, каждая из которых равна третьей, также равны между собой. Это означает, что справедливо и следующее:

$$P \text{ (S} \mid T) \text{ } P \text{ } (T) = P \text{ } (T \mid S) \text{ } P \text{ } (S)$$

Это безобидное с виду уравнение стало известно как «правило Байеса». Если посмотреть на него внимательнее, то обнаружится, что оно предлагает общее решение для проблемы обратной вероятности. Оно говорит: если мы знаем вероятность S при T, P ($S \mid T$), то мы сможем вычислить вероятность T при S, P ($T \mid S$) — конечно, при условии, что P (T) и P (S) нам известны. Это, пожалуй, самая важная функция правила Байеса в статистике: мы можем напрямую оценить условную вероятность в одном направлении, где наше суждение надежнее, и применить математику, чтобы получить условную вероятность в другом направлении, для которого наше суждение довольно туманно. Уравнение тоже играет эту роль в байесовских сетях; мы сообщаем компьютеру прямые вероятности, а компьютер выдает обратные вероятности, когда это необходимо.

Чтобы увидеть, как правило Байеса действует в примере с чайной, предположим, что вы не потрудились вычислить P ($T \mid S$) и оставили таблицу с данными дома. Однако вы

почему-то помните, что половина из заказавших чай также заказала пирожные. Тут ваш босс задает неожиданный вопрос: «Какая доля заказавших пирожные также заказала и чай?» Нет повода для паники — вы можете вычислить это на основании иных вероятностей. Правило Байеса говорит, что $P(T \mid S)$ $(5/12) = (1/2)(2/3)$, поэтому ваш ответ — $P(T \mid S) = 4/5$, потому что $4/5$ — единственное значение для $P(T \mid S)$, которое сделает уравнение верным.

Также мы можем посмотреть на правило Байеса как на способ по-новому оценить нашу веру в определенную гипотезу. Это чрезвычайно важно понимать, потому что человеческие представления о событиях в будущем во многом опираются на частоту похожих событий в прошлом. Например, когда клиентка заходит в кафе, мы, ориентируясь на поведение похожих клиенток в прошлом, думаем, что, вероятно, она закажет чай. Но, если она сначала попросит пирожное, наша уверенность даже возрастет. Более того, возможно, мы предложим: «И чаю к пирожным?» Правило Байеса просто позволяет нам подкрепить эти рассуждения цифрами. Из табл. 1 видно, что предыдущая вероятность заказа чая (когда клиентка только вошла и еще ничего не заказала) равна $2/3$. Но если клиентка заказывает пирожные, у нас появляется дополнительная информация о ней, которой не было раньше. В этом случае вероятность заказа чая (когда уже заказаны пирожные) выглядит так: $P(T \mid S) = 4/5$.

С математической точки зрения в этом и состоит правило Байеса. Оно кажется почти банальным. Здесь нет ничего, кроме понятия условной вероятности и небольшой дозы древнегреческой логики. Вы можете задать оправданный вопрос: как такая небольшая «фишка» сделала Байеса известным и почему люди спорили о ней 250 лет. В конце концов, математические факты должны разрешать противоречия, а не создавать их.

Здесь я должен признаться, что в примере с чайной, выводя правило Байеса из полученных данных, я опустил два весьма существенных возражения — одно философское и одно практическое. Философское возражение происходит из интерпретации вероятностей как степени веры, которую мы подспудно исполь-

зовали в случае с чайной. Кто вообще сказал, что убеждения действуют или должны действовать как пропорциональные отношения данных?

Загвоздка в этом философском споре состоит в том, можно ли полноценно перевести выражение «при том, что я знаю» на язык вероятностей. Даже если мы согласимся, что безусловные вероятности вроде $P(S)$, $P(T)$ и $P(S \text{ and } T)$ отражают мою степень убежденности в этих предложениях, кто может сказать, что если оценить степень моей веры в T, она будет равна отношению $P(S \text{ and } T) / P(T)$, как утверждает правило Байеса? Будет ли «при том, что известно T» одним и тем же во всех случаях, где встретилось T? Язык вероятностей, выраженный в таких символах как $P(S)$, создавался, чтобы выразить понятие частоты в азартных играх. Но выражение «при том, что известно» — эпистемологическое и должно управляться логикой знания, а не логикой частоты и пропорций.

С философской точки зрения достижение Томаса Байеса состоит в том, что он предложил формальное определение условной вероятности как $P(S \mid T) = P(S \text{ and } T) / P(T)$. По общему признанию, его эссе имеет довольно размытые формулировки; у него нет термина для условной вероятности, и вместо него он использует громоздкий оборот «вероятность второго [события] в условиях предположения, что первое произойдет». Только в 1880-х годах было признано, что отношение «при условии, что» заслуживает собственный символ, и только в 1931 году Харолд Джефрис (более известный как геофизик, чем как теоретик вероятности) ввел стандартную сегодня вертикальную черту в $P(S \mid T)$.

Как мы видели, правило Байеса с формальной точки зрения — элементарное следствие его определения условной вероятности. Но с эпистемологической точки зрения оно далеко не элементарно. Более того, оно действует как нормативное правило для регуляции убеждений в ответ на доказательства. Другими словами, байесовское правило стоит рассматривать не только как удобное определение для нового понятия условной вероятности, но как попытка на практике достоверно представить английское выражение «при условии, что я знаю».

Помимо прочего, оно означает, что вера в *S*, которую приобретает человек, открыв *T*, всегда не менее сильна, чем вера, которую человек питает по отношению к *S* и *T* до того, как откроет *T*. Более того, оно подразумевает, что чем удивительнее факт *T*, т.е. чем меньше *P* (*T*), тем сильнее должна быть вера в его причину *S*. Не случайно Байес и его друг Прайс, будучи епископальными священниками, видели в этом удачную отповедь Юму. Если *T* — чудо («Христос воскрес из мертвых»), а *S* — тесно связанная с ним гипотеза («Христос — сын Бога»), наша степень веры в *S* радикально повышается, когда мы точно знаем, что *T* — правда. Чем чудеснее чудо, тем больше доверия заслуживает гипотеза, которая обосновывает его возникновение. Это объясняет, почему авторы Нового Завета были так сильно впечатлены свидетельствами очевидцев.

А теперь я хотел бы обсудить практическое возражение правилу Байеса, которое, возможно, становится важнее, когда мы выходим из рамок теологии и переходим на территорию науки. Если попытаться применить это правило к головоломке с бильярдным шаром, чтобы найти $P(L \mid x)$, то понадобится величина физики бильярдных шаров, недоступная нам: нам нужна априорная вероятность длины L, которую так же сложно определить, как и желаемую $P(L \mid x)$. Более того, эта вероятность будет значительно отличаться в зависимости от индивидуального опыта каждого со столами разной длины. Человек, который никогда в жизни не видел стола для снукера, будет сильно сомневаться в том, что *L* может оказаться больше 10 футов. Однако человек, который видел только столы для снукера и не видел классического бильярдного стола, счел бы *L* меньше 10 футов крайне маловероятной. Эту переменчивость, также известную как субъективность, иногда считают недостатком причинного вывода по Байесу. Между тем есть мнение, что она дает мощное преимущество, поскольку позволяет выразить личный опыт математически и объединить его с данными — последовательно и прозрачно. Правило Байеса направляет наши рассуждения в тех случаях, когда подводит обычная интуиция или вмешиваются эмоции. Мы продемонстрируем это преимущество на примере знакомой всем нам ситуации.

Предположим, вы прошли медицинское обследование, чтобы узнать, есть ли у вас заболевание, и результат оказался положительным. Насколько вероятно, что вы действительно больны? Ради конкретности предположим, что речь идет о раке груди, а метод обследования — маммография. Здесь *прямая* вероятность — это вероятность положительного результата в случае, если вы действительно больны: P (*обследование* | *болезнь*). Врач назвал бы это «чувствительностью» обследования, подразумевая его способность правильно выявлять болезнь. Как правило, это одинаковая величина для всех пациентов, потому что она зависит только от технических возможностей прибора, выявляющего связанные с заболеванием отклонения. *Обратная* вероятность, скорее всего, окажется для вас более важной: какова вероятность, что вы больны, если результат оказался положительным? Это P (*болезнь* | *обследование*), и здесь информация идет не в причинном направлении, а от результата обследования к вероятности болезни. Вероятность не одинакова для всех типов пациентов; безусловно, положительный результат будет более тревожным для пациентки с семейным анамнезом болезни, чем для пациентки без такого анамнеза.

Обратите внимание, что мы начали говорить о причинных и непричинных направлениях. Мы не сделали этого в примере с чайной, потому что там было не важно, что делали в первую очередь — заказывали чай или просили пирожные. Было важно одно: какую условную вероятность можно оценить. Но причинно-следственный контекст проясняет, почему мы чувствуем себя менее уверенно, оценивая обратную вероятность, а в эссе Байеса прямо говорится, что его интересовала именно эта задача.

Предположим, 40-летней женщине сделали маммографию, чтобы проверить, нет ли у нее рака груди, и результаты оказались положительными. Гипотеза D (от англ. *disease* — «болезнь») состоит в том, что у нее рак. Доказательство, T (от англ. *test* — «анализ, обследование») представляет собой результат маммографии. Насколько стоит верить этой гипотезе? Следует ли делать операцию?

Мы можем ответить на эти вопросы, переписав правило Байеса следующим образом:

$$\text{Обновленная вероятность } D = P(D \mid T) = \\ = \text{Отношение правдоподобия} \times \\ \times \text{ Априорная вероятность } D \quad (1),$$

где новый термин «отношение правдоподобия» определяется как $P(T \mid D) / P(T)$. Он измеряет, насколько вероятнее положительный результат обследования у людей с этим заболеванием, чем у населения в целом. Таким образом, уравнение (1) говорит, что новые данные T увеличивают вероятность D на фиксированную пропорцию независимо от того, какой была априорная вероятность.

Приведем пример, чтобы увидеть, как работает эта важная концепция. У обычной 40-летней женщины вероятность заболеть раком груди в следующем году — приблизительно 1:700, поэтому мы будем использовать ее в качестве априорной вероятности.

Чтобы вычислить отношение правдоподобия, нам нужно знать $P(T \mid D)$ и $P(T)$. В медицинском контексте $P(T \mid D)$ — это чувствительность маммограммы, т.е. вероятность положительного результата, если у пациентки рак. По данным Консорциума по надзору за раком груди (*B*reast Cancer S*U*rveillance Consorti*U*m; *B*CSC), чувствительность маммограммы для 40-летних женщин составляет 73%.

Со знаменателем $P(T)$ дело обстоит немного сложнее. Положительный результат T может быть получен как от пациенток, у которых есть эта болезнь, так и от пациенток, у которых ее нет. Таким образом, $P(T)$ должно быть средневзвешенным значением $P(T \mid D)$ (вероятность положительного результата у тех, кто болеет) и $P(T \mid \sim D)$ (вероятность положительного результата у тех, кто этим не болеет). Второй называют уровнем ложноположительных результатов. Согласно *B*CSC, уровень ложноположительных результатов для 40-летних женщин составляет около 12%.

Почему средневзвешенная? Потому что здоровых женщин ($\sim D$) намного больше, чем женщин, больных раком (D). Фактически только 1 из 700 женщин страдает этим недугом, а остальные 699 — нет, поэтому вероятность положительного результата теста для случайно выбранной женщины должна гораздо сильнее зависеть от 699 женщин, у которых нет рака, чем от одной женщины, у которой он есть.

Получить средневзвешенное значение можно с помощью следующих вычислений: $P(T) = 1/700 \cdot 73\% + 699/700 \cdot 12\%$ а 12,1%. Коэффициенты обусловлены тем, что только у 1 из 700 женщин вероятность положительного результата составляет 73%, а у остальных 699—12%. Как и следовало ожидать, $P(T)$ оказался очень близок к уровню ложноположительных результатов.

Теперь, когда мы знаем $P(T)$, наконец-то можно вычислить обновленную вероятность — шанс женщины заболеть раком груди после того, как результат окажется положительным. Отношение правдоподобия составляет 73% / 12,1% ≈ 6. Как я уже говорил, это фактор, на который мы увеличиваем ее априорную вероятность, чтобы вычислить обновленную вероятность рака. Поскольку ее априорная вероятность была равна 1 из 700, ее обновленная вероятность составляет $6 \cdot 1/700$ а 1/116. Другими словами, у нее все еще есть вероятность заболеть раком и она составляет менее 1%.

Вывод поразительный. Я думаю, большинство 40-летних женщин с положительным результатом маммографии были бы изумлены, узнав, что шанс заболеть раком груди у них составляет менее 1%. Рис. 14 поможет понять причины: крошечное число истинно положительных результатов (т.е. женщин с раком груди) несоизмеримо с огромным числом ложноположительных результатов. Наше удивление по поводу этого явления объясняется общей когнитивной путаницей между прямой вероятностью, которая хорошо изучена и тщательно задокументирована, и обратной вероятностью, необходимой для принятия личного решения.

Конфликт между нашим восприятием и реальностью частично объясняет протесты, возникшие, когда рабочая группа

по профилактике болезней (*Preventive Services Task Force*) в США в 2009 году рекомендовала 40-летним женщинам не проходить ежегодную маммографию. Рабочая группа понимала то, чего не осознавали многие женщины: положительный результат обследования в этом возрасте с большей вероятностью будет ложной тревогой и многие женщины в таких случаях испугаются зря (и получат ненужное лечение).

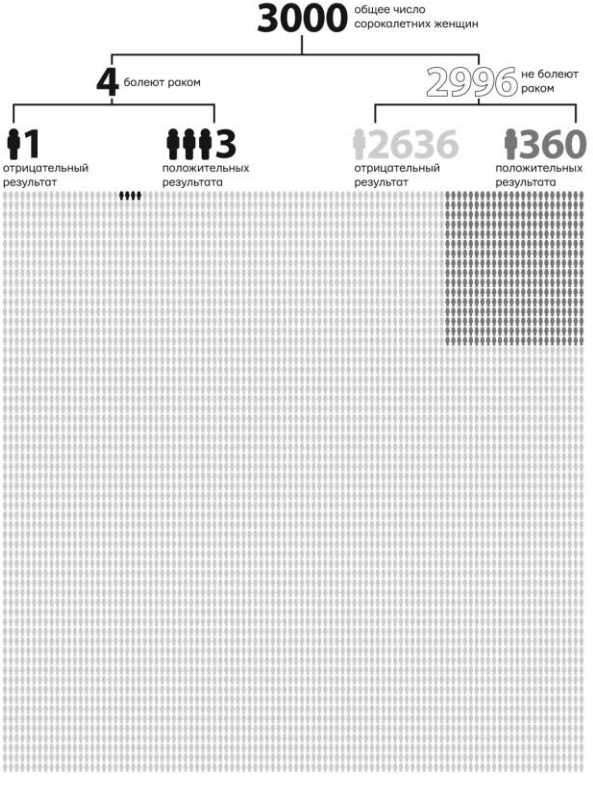


Рис. 14. В этом примере, основанном на количестве ложноположительных и ложноотрицательных результатов, предоставленных Консорциумом по надзору за раком молочной железы, только 3 из 363 40-летних женщин с положительным результатом обследования на рак груди действительно оказались больны (пропорции не совсем соответствуют тексту из-за округления) (источник: инфографика Маян Харел)

Но все было бы иначе, если бы у нашей пациентки был ген, который подвергал бы ее высокому риску рака груди, скажем с одним шансом из 20 в течение следующего года. Тогда положительный результат повысил бы вероятность почти до одного из трех. Для женщины в этой ситуации шансы, что обследование даст жизненно важную информацию, намного выше. Вот почему рабочая группа рекомендует женщинам из группы высокого риска делать маммограммы ежегодно.

Этот пример показывает, что P (*болезнь* | *обследование*) неодинаков для всех; вероятность зависит от контекста. Если вы знаете, что изначально подвержены высокому риску заболевания, правило Байеса позволяет вам учесть эту информацию. Или, если вы знаете, что риска нет, обследование просто не нужно. Напротив, P (*обследование* | *болезнь*) не зависит от того, находитесь вы в группе риска или нет. Вероятность устойчива к таким вариациям, что до некоторой степени объясняет, почему врачи систематизируют и передают свои знания с помощью прямых вероятностей. Вариации — это свойства самой болезни, ее стадии развития или чувствительности детекторов; следовательно, они остаются относительно инвариантными к причинам заболевания (эпидемия, диета, гигиена, социально-экономический статус, семейный анамнез). Обратная вероятность P (*болезнь* | *обследование*) чувствительна к этим условиям.

Читатель, интересующийся историей, наверняка задастся вопросом, как Байес справился с субъективностью P (L), где L — длина бильярдного стола. Ответ состоит из двух частей. Во-первых, Байеса интересовала не длина стола как таковая, а связанные с ней последствия (т.е. вероятность, что следующий шар окажется в каком-то определенном месте на столе). Во-вторых, Байес предположил, что L определяется механически, когда бильярдный шар отправляют с большего расстояния, скажем, L^*. Таким образом, он наделил P (L) объективностью и преобразовал задачу так, что априорные вероятности можно оценить на основе данных, как мы видим в образцах с чайной и маммограммой.

Во многих отношениях правило Байеса — квинтэссенция научного метода. Описание последнего в учебнике выглядит примерно так: 1) сформулируйте гипотезу; 2) выведите проверяемое следствие гипотезы; 3) проведите эксперимент и соберите доказательства и 4) пересмотрите веру в гипотезу. Обычно учебники разбирают простые тесты типа «да или нет» и полученные результаты; доказательства либо подтверждают, либо опровергают гипотезу. Но жизнь и наука не бывают такими простыми! Все полученные данные отличаются некоторой неопределенностью. И правило Байеса показывает нам, как выполнить шаг 4 в реальном мире.

От байесовского правила к байесовским сетям

В начале 1980-х проектирование искусственного интеллекта зашло в тупик. С тех пор как Алан Тьюринг впервые изложил задачу в статье 1950 года «Вычислительные машины и интеллект», ведущим подходом в этой области были так называемые системы на основе правил или экспертные системы, которые организуют человеческое знание как набор конкретных и общих фактов и используют правила логического вывода, чтобы связать их. Например: Сократ — человек (конкретный факт). Все люди смертны (общий факт). Из этой базы знаний мы (или разумная машина) можем вывести тот факт, что Сократ смертен, используя универсальное правило логического вывода: если все *A* являются *B* и *x* является *A*, то *x* является *B*.

Теоретически это был годный подход, но жесткие правила вряд ли могут отразить знания из реальной жизни. На деле мы все время сталкиваемся с исключениями из правил и неопределенностями в данных, даже когда этого не осознаем. К 1980 году стало ясно, что экспертным системам трудно делать правильные выводы из неопределенных знаний. Компьютер не мог воспроизвести процесс, с помощью которого человек-специалист приходит к логическому выводу, потому

что сами специалисты не могли выразить свой мыслительный процесс на языке, доступном системе.

Таким образом, конец 1970-х был временем брожения умов: сообщество исследователей ИИ пыталось найти способ справиться с неопределенностью. В идеях недостатка не было. Лотфи Заде из Калифорнийского университета в Беркли предложил «нечеткую логику», в которой утверждения, не являясь ни истинными, ни ложными, принимают ряд возможных значений истинности. Гленн Шейфер из Канзасского университета предложил «функции убеждений», которые приписывают каждому факту две вероятности: одна указывает, насколько вероятно, что он «возможен», другая — насколько вероятно, что он «доказуем». Эдвард Фейгенбаум и его коллеги из Стэнфордского университета попробовали работать с «факторами достоверности», добавив числовые меры неопределенности в детерминистские правила логического вывода.

К сожалению, несмотря на всю изобретательность, эти подходы имели общий недостаток: они моделировали эксперта, а не мир и поэтому нередко давали непредвиденные результаты. Например, они не могли работать одновременно в диагностическом и прогностическом режимах, что является бесспорным преимуществом правила Байеса. При подходе, основанном на факторе определенности, правило «Если огонь, то дым (с определенностью c_1)» не может согласованно сочетаться с утверждением «Если дым, то огонь (с определенностью c_2)», не вызывая бесконтрольного роста уверенности.

В то время также рассматривался подход, основанный на вероятностях, однако он сразу приобрел дурную славу из-за огромных потребностей в памяти для хранения и очень долгого времени обработки. Я вышел на арену довольно поздно, в 1982 году, с очевидным, но радикальным предложением: вместо того чтобы заново изобретать теорию неопределенности с нуля, оставим вероятность в качестве защитницы здравого смысла и просто исправим ее недостатки в вычислительном плане. А именно, вместо того чтобы представлять вероятность в огромных таблицах, как это делали раньше, выразим ее в виде сети слабо связанных переменных. Если мы разрешим каждой

переменной взаимодействовать только с несколькими соседними, это позволит преодолеть вычислительные препятствия, которые помешали другим исследователям вероятностей.

Эта идея пришла ко мне не во сне; она почерпнута из статьи Дэвида Румельхарта, когнитивиста из Калифорнийского университета в Сан-Диего и пионера в проектировании нейросетей. Его статья о детском чтении, опубликованная в 1976 году, показала, что это сложный процесс, в ходе которого нейроны на многих разных уровнях действуют одновременно (рис. 15). Одни нейроны распознают отдельные особенности — круги или линии. Над ними другой слой нейронов объединяет эти формы и строит предположения о том, что за буква получается из них.

На рис. 15 показано, как нейросеть борется с большой долей неопределенности применительно ко второму слову. На уровне букв это может быть FHP, но на уровне слов такое сочетание не имеет особого смысла. Здесь предположительны FAR, CAR или FAT. Нейроны переводят информацию на синтаксический уровень, который определяет, что после слова THE ожидается существительное. Наконец, эта информация полностью передается на семантический уровень, где учитывается, что в предыдущем предложении упоминался Volks*W*agen, а значит, искомым сочетанием будет THE CAR (ЭТА МАШИНА), относящееся к тому самому Volks*wa*gen. Важнее всего здесь, что нейроны передают информацию туда и обратно, сверху вниз, снизу вверх и из стороны в сторону. Это система со многими параллельными процессами, которая сильно отличается от нашего представления о мозге как о монолитной системе с централизованным управлением.

Читая статью Румельхарта, я убеждался в том, что любой искусственный интеллект должен будет моделировать себя на основе наших знаний о нейронной обработке информации у человека и что машинное мышление в условиях неопределенности должно использовать похожую архитектуру передачи сообщений. Но что же это за сообщения? На понимание этого у меня ушел не один месяц. И наконец я осознал, что эти сообщения были условными вероятностями в одном направлении и отношениями правдоподобия в другом.

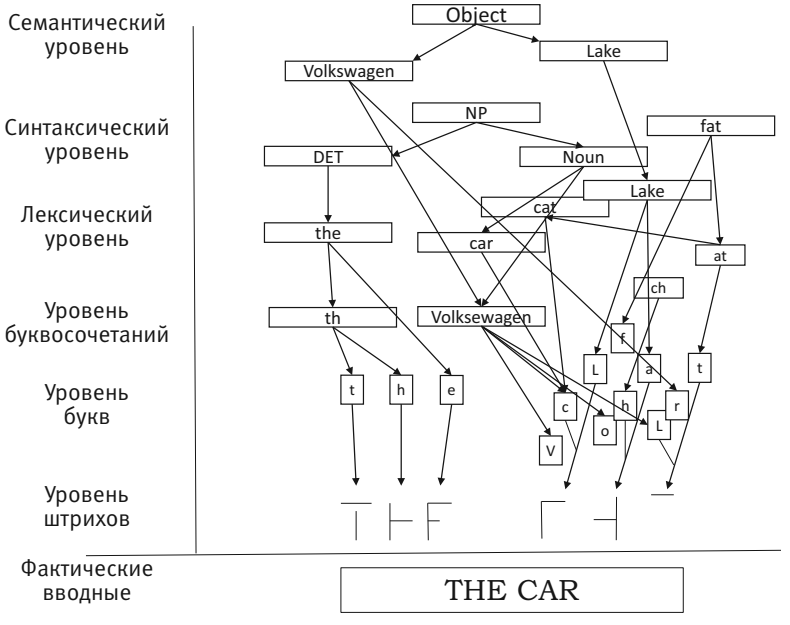


Рис. 15. Набросок Дэвида Румельхарта показывает, как сеть передачи сообщений учится читать сочетание THE CAR (источник: любезно предоставлено Центром исследований мозга и познания Калифорнийского университета в Сан-Диего)

Точнее, я предполагал, что сеть будет иерархической — со стрелками, ведущими от верхних нейронов к нижним или от «родительских узлов» к «дочерним узлам». Каждый узел будет отправлять всем соседям (как выше, так и ниже в иерархии) сообщение о своей текущей степени уверенности в переменной, которую отслеживает (например, «Я на две трети уверен, что эта буква — R»). Получатель будет обрабатывать сообщение двумя разными способами, в зависимости от его направления. Если сообщение идет от «родителя» к «ребенку», то «ребенок» обновит степень уверенности, используя условные вероятности, подобные тем, которые мы видели в образце с чайной. Если сообщение передается от «ребенка» к «родителю», то родитель обновит свою степень уверенности, умножив их на отношение правдоподобия, как в случае с маммограммой.

Повторное применение этих двух правил к каждому узлу в сети называется распространением степени уверенности. В ретроспективе видно, что в этих правилах нет ничего произвольного или выдуманного; они находятся в строгом соответствии с правилом Байеса. Настоящий вызов состоял в том, чтобы гарантировать удобное равновесие в конце — независимо от того, в каком порядке отправляются эти сообщения; более того, окончательное равновесие должно представлять «правильное» отражение веры в переменные. Под «правильным» я имею в виду такой же результат, как если бы мы проводили вычисления с помощью методов из учебника, а не путем передачи сообщений.

Это задача заняла меня и моих студентов, а также моих коллег на несколько лет. Но к концу 1980-х годов мы преуспели до такой степени, что байесовские сети стали практической схемой машинного обучения. За следующие 10 лет сфера их применения, например, для фильтрации спама и распознавания голоса, постоянно расширялась. Однако к тому времени я уже пытался подняться по Лестнице Причинности, передав вероятностную сторону байесовских сетей в другие надежные руки.

Байесовские сети: что причины говорят о данных

Хотя Байес этого не знал, его правило обратной вероятности представляет собой простейшую байесовскую сеть. Мы уже видели ее в нескольких обличиях: чай → пирожные, болезнь → анализ и, в более общем контексте, гипотеза → подтверждения. В отличие от диаграмм причинности, с которыми мы будем иметь дело в течение всей книги, байесовские сети не подразумевают, что стрелки обозначают причинно-следственные связи. Стрелка просто значит, что нам известна «прямая» вероятность: P (*пирожные* | *чай*) or P (*тест* | *болезнь*). Правило Байеса показывает нам, как развернуть процедуру обратно, в частности, путем умножения априорной вероятности на отношение правдоподобия.

Формально распространение уверенности осуществляется абсолютно одинаково, и неважно, обозначают ли стрелки причинно-следственные связи. Тем не менее у вас может появиться интуитивное ощущение, что во втором случае мы сделали нечто более осмысленное. Это потому, что наши мозги оснащены специальным аппаратом для понимания причинно-следственных связей (например, между раком и маммографией). Для чистых ассоциаций (скажем, между чаем и пирожными) это не работает.

Следующий этап после сети из двух узлов с одной связью — конечно же, сеть из трех узлов с двумя связями, которую я буду называть связкой. Это строительные блоки во всех байесовских (и причинно-следственных) сетях. Существуют три основных типа связок, с помощью которых мы можем описать любое использование стрелок в сети.

1. **A → B → C**. Эта связка — самый простой образец цепочки или медиации. В науке *B* часто считают механизмом или посредником, который передает действие *A* на *C*. Знакомый пример — *огонь → дым → тревога*. Хотя мы называем это «пожарной сигнализацией», на самом деле она реагирует на дым. Огонь как таковой не запускает сигнализацию, поэтому стрелки между огнем и тревогой нет. Также огонь не запускает сигнализацию ни через какую другую переменную вроде температуры. Сигнализация реагирует только на молекулы дыма в воздухе. Если отменить это звено в цепочке, скажем отсосав все молекулы дыма с помощью вытяжки, то тревоги не будет.

Это наблюдение приводит к важному концептуальному выводу о цепочках: посредник *B* «отсеивает» информацию об *A*, не давая ей доступа к *C*, и наоборот (впервые на это указал Ханс Рейхенбах, немецко-американский философ науки). Так, если мы уже знаем о присутствии или отсутствии дыма, информация об огне не может дать нам оснований, чтобы в большей или меньшей степени верить сигнализации. Эта стабильность веры — понятие первого уровня; следовательно, можно ожидать, что мы будем наблюдать его и в данных, если они доступны. Предположим, у нас есть база данных обо всех случаях возгорания, дыма или срабатывания сигнализа-

ции. Если бы мы смотрели только на те строки, где *дым* = 1, то ожидали бы, что *тревога* = 1 всякий раз, независимо от того, *огонь* = 0 или *огонь* = 1. Этот эффект отсеивания действует, если следствие не является детерминированным. Представьте себе неисправную систему сигнализации, которая не срабатывает правильно в 5% случаев. Если посмотреть только на строки, где *дым* = 1, окажется: вероятность, что *тревога* = 1, одинакова (95%), и неважно, *огонь* = 0 или *огонь* = 1.

Просмотр только тех строк в таблице, где *дым* = 1, называется ограничением по переменной. Подобным образом мы говорим, что огонь и тревога ограниченно независимы, учитывая значение дыма. Это важно знать, если вы программируете машину, чтобы обновить ее убеждения; ограниченная независимость дает машине право сосредоточиться на значимой информации и игнорировать всю остальную. Всем нам необходимо такое право для повседневной мыслительной деятельности, иначе мы будем постоянно гоняться за ложными сигналами. Но как же решить, какую информацию игнорировать, если каждый новый ее фрагмент меняет границу между значимым и неважным? К людям это понимание приходит естественным путем. Даже трехлетние малыши понимают эффект отсеивания, хотя у них нет для него названия. Их инстинкт, вероятно, основан на некой репрезентации в уме, возможно напоминающей причинную диаграмму. Но у машин нет такого инстинкта, и это одно из обстоятельств, по которым мы снабжаем их причинными диаграммами.

2. **A** ← **B** → **C**. Этот тип связки называется «вилка», и *B* часто считают общей причиной или общим осложнителем для *A* и *C*. Осложняющая переменная обеспечивает статистическую корреляцию между *A* и *C*, хотя между ними нет прямой причинной связи. Вот хороший пример (от Дэвида Фридмана): *размер обуви* ← *возраст ребенка* → *навыки чтения*. Дети, у которых больше размер обуви, обычно лучше читают. Но это не причинно-следственные отношения. Если дать ребенку обувь большего размера, он не станет от этого лучше читать! Напротив, обе переменных объясняются третьей — возрастом

ребенка. У более старших детей обувь большего размера, и одновременно они более продвинутые читатели.

Мы можем избавиться от этой ложной корреляции, как называли ее Карл Пирсон и Джордж Удни Юл, ограничив нашу выборку возрастом ребенка. Так, если взять только семилетних детей, мы не будем ожидать какой-либо зависимости между размером обуви и умением читать. Как и в случае с цепочкой, *A* и *C* условно независимы, если дано *B*.

Прежде чем перейти к третьей связке, необходимо кое-что прояснить. Условная независимость, которую я только что упомянул, проявляется всякий раз, когда мы смотрим на эти связки в изоляции. Если их окружают дополнительные причинные связи, последние необходимо принять во внимание. Чудо байесовских сетей состоит в том факте, что три вида связок, которые мы описываем в изоляции, достаточны, чтобы увидеть любую независимость, подразумеваемую байесовской сетью, какой бы сложной она ни была.

3. **A → B ← C.** Это самая интересная связка под названием «коллайдер». Феликс Элверт и Крис Уиншип проиллюстрировали ее, используя три характеристики голливудских актеров: *талант → известность ← красота*. Здесь мы утверждаем, что и талант, и красота способствуют успеху актера, но красота и талант совершенно не связаны друг с другом у людей в целом.

Сейчас мы увидим, что принцип коллайдера работает совершенно противоположно цепочке или вилке, если мы ограничим значение переменной в середине. Если *A* и *C* независимы с самого начала, ограничение по *B* сделает их зависимыми. Например, если мы посмотрим только на известных актеров и актрис (другими словами, мы наблюдаем переменную *известность* =1), то мы увидим негативную корреляцию между талантом и красотой: обнаружив, что актер или актриса не обладает красотой, мы укрепляемся в убеждении, что он или она отличается талантом.

Эту негативную корреляцию порой называют ошибкой коллайдера или эффектом достаточного объяснения. Для про-

стоты представим, что для статуса звезды не нужны ни талант, ни красота — достаточно чего-то одного. Тогда, если актер *A* особенно хорош, это «достаточно объясняет» его успех и ему не нужно быть красивее среднего человека. В свою очередь, если актер *B* особенно плох, то единственный способ объяснить его успех — привлекательная внешность, т.е. с учетом результата *известность* = 1 талант и красота связаны обратно, даже если они не связаны между собой у людей в целом. Но и в более реалистичной ситуации, где успех — сложная функция, зависящая от красоты и таланта, эффект достаточного объяснения все же присутствует. Однако этот образец несколько апокрифичен, потому что красоту и талант трудно измерить объективно; тем не менее ошибка коллайдера вполне реальна и в этой книге мы увидим множество тому примеров.

Эти три связки — цепи, вилки и коллайдеры — подобны замочным скважинам в двери, разделяющей первый и второй уровни Лестницы Причинности. Заглянув в них, мы можем увидеть секреты причинного процесса, который породил наблюдаемые нами данные. Каждая символизирует определенный принцип причинно-следственной связи и оставляет след в виде зависимости и независимости данных друг от друга при определенных условиях. В публичных лекциях я часто называю их дарами богов, поскольку они позволяют тестировать причинно-следственную модель, открывать новые модели, оценивать эффекты интервенции и многое другое. Тем не менее, взятые в отдельности, они позволяют лишь мельком взглянуть на ситуацию. Нам нужен ключ, который полностью откроет дверь и позволит выйти на второй уровень. Этот ключ, о котором мы узнаем из главы 7, включает все три связки и называется *d*-разделением. Его концепция позволяет нам увидеть, какого рода зависимости можно ожидать в данных при разных шаблонах и путях в модели причинно-следственных связей. Такая фундаментальная связь между причинами и вероятностями составляет основной вклад байесовских сетей в науку о причинном выводе.

Где мой чемодан? От Ахена до Занзибара

Пока я сделал акцент только на одном аспекте байесовских сетей, а именно на диаграмме и стрелках, которые в идеале ведут от причины к следствию. В самом деле, эта диаграмма — двигатель байесовской сети. Но для любого двигателя требуется топливо. В данном случае это *таблица условных вероятностей*.

По-другому это можно выразить так: диаграмма описывает отношение вероятностей в качественном виде, но если нужны количественные ответы, то необходимы и количественные вводные. В байесовской сети нужно определить условную вероятность каждого узла с учетом его «родителей» (вспомним, что «родительские узлы» ведут к «дочерним»). Это прямые вероятности, P (*подтверждение* | *гипотеза*).

В случае когда A — корневой узел и на него не указывают стрелки, надо просто определить априорную вероятность для каждого состояния A. В нашей второй сети *болезнь* (D) → *обследование* (T) D — корневой узел. Таким образом, мы определили априорную вероятность того, что пациентка больна (1/700 в нашем примере), и того, что она не больна (699/700 в нашем примере).

Описывая A как корневой узел, мы на самом деле не подразумеваем, что у A нет предшествующих причин. Вряд ли какая-то переменная имеет право на такой статус. На самом деле мы имеем в виду, что любые предыдущие причины A могут быть адекватно обобщены в априорной вероятности P (A) *т*ого, что А верно. Так, в случае с болезнью и обследованием семейный анамнез может быть причиной заболевания. Но до тех пор, пока мы уверены, что семейный анамнез не повлияет на переменную *обследование* (как только мы узнаем статус *болезни*), нет необходимости представлять ее как узел на графике. Однако, если существует причина заболевания, которая также напрямую влияет на *обследование*, то эта причина должна быть явно представлена на диаграмме.

В случае если у A есть родитель, она должна «послушать» его, прежде чем определиться с собственным состоянием. В примере с маммографией родителем *обследования* (T) была *болезнь* (D).

Мы можем показать этот процесс «слушания» в таблице 2 × 2 (табл. 2). Скажем, если *T* «слышит», что *D* = 0, то в 88% случаев *T* будет равно 0 (*T*=0), в 12% — 1 (*T* = 1). Обратите внимание на то, что во второй части таблицы содержится та же информация, которую предоставил Консорциум по надзору за раком груди: доля ложноположительных результатов (правый верхний угол) — 12%, а чувствительность — 73%. Значения в двух оставшихся клетках дополняют сумму до 100%.

Таблица 2. Простая таблица условной вероятности

Вероятность → при ↓	***T= 0***	***T =1***
D = 0	88	12
D = 1	27	73

По мере того как мы переходим к более сложным сетям, таблица условной вероятности тоже становится сложнее. Скажем, если у нас есть узел с двумя родителями, в таблице условной вероятности необходимо учитывать четыре возможных состояния обоих родителей. Давайте разберем конкретный пример, который предложили Стефан Конради и Лайонел Джофф из BayesiaLab, Inc. Это сценарий, знакомый всем путешественникам. Мы назовем его «Где мой чемодан?».

Предположим, вы только что приземлились в Занзибаре, сделав очень быструю пересадку в Ахене, и ждете, пока ваш чемодан появится на багажной карусели. Другие пассажиры уже получают багаж, но вы все ждете… ждете… и ждете… Каковы шансы на то, что ваш чемодан действительно сделал пересадку в Ахене на рейс до Занзибара? Ответ зависит, конечно, от того, сколько вы уже ждете. Если сумки только появились на ленте, возможно, стоит потерпеть и подождать еще. Но если прошло много времени, перспективы ухудшаются. Мы выразим повод для переживаний количественно, сделав диаграмму причинности (рис. 16).

Рис. 16. Диаграмма причинности для примера с чемоданом в аэропорту

Эта диаграмма иллюстрирует интуитивную идею о том, что у появления чемодана на ленте багажной карусели есть две причины. Для начала он должен находиться в самолете — в противном случае он точно не появится на ленте. Во-вторых, присутствие чемодана на ленте становится менее вероятным с течением времени, если он вообще был на борту…

Чтобы превратить диаграмму причинности в байесовскую сеть, надо определиться с таблицами условной вероятности. Скажем, все чемоданы в аэропорту Занзибара разгружаются в течение 10 минут. (В Занзибаре все очень эффективны!) Предположим также, что вероятность успешной пересадки вашего чемодана *P* (*чемодан в самолете* = верно) равна 50%. (Прошу прощения, если это заденет кого-то из сотрудников ахенского аэропорта. Я всего лишь использую пример Конради и Джоффа. Сам я предположил бы более высокую вероятность — 95%).

Настоящая рабочая лошадка этой байесовской сети — таблица условной вероятности для чемодана на ленте багажной карусели (табл. 3).

Хотя это довольно большая таблица, понять ее должно быть легко. Первые 11 строк говорят о том, что если чемодан не попал в самолет (*чемодан в самолете* = неверно), то, сколько бы ни прошло времени, он не окажется на ленте багажной карусели (*лента* = неверно), т.е. *P* (*лента* = неверно | *чемодан в самолете* = неверно) равна 100%. Это объясняет 100 в первых 11 строках.

Другие 11 рядов говорят, что чемоданы выгружаются с самолета с устойчивой скоростью. Если ваш чемодан правда в самолете, есть 10%-ная вероятность, что его выгрузят

в первую минуту, 10%-ная вероятность для второй минуты и т.д. Так, через 5 минут вероятность, что его выгрузили, будет равна 50%, поэтому мы видим 50 P (*лента* = верно | *чемодан в самолете* = верно, *время* = 5). Через 10 минут все чемоданы выгружены, так что P (лента = верно | *чемодан в самолете* = верно, *время* = 10) равна 100%. Таким образом, в последней клетке таблицы 100.

Самое интересное, что можно сделать с этой байесовской сетью, как и с большинством байесовских сетей, — решить проблему обратной вероятности. Если прошло *x* минут и я до сих пор не получил чемодан, какова вероятность того, что он на самолете? Правило Байеса автоматизирует это вычисление и показывает интересный момент. Через минуту эта вероятность еще равно 47% (вспомним, что нашим изначальным предположением была вероятность 50%). Через 5 минут вероятность снижается до 33%. Через 10 минут, конечно же, она падает до нуля. Рис. 17 показывает, как вероятность распределяется во времени, и это можно назвать «кривой расставания с надеждой». Мне интересно, что это *правда* кривая: думаю, большинство людей ожидают увидеть здесь прямую линию. Вообще, отсюда следует довольно оптимистичный вывод: не отчаивайтесь слишком рано! Кривая показывает, что, когда проходит половина отведенного времени, стоит расстаться всего лишь с третью надежды.

Таблица 3. Более сложная таблица условной вероятности

Вероятность → при ↓		**Лента = неверно**	**Лента = верно**
Чемодан в самолете	**Истекшее время**		
Неверно	0	100	0
Неверно	1	100	0

Окончание таблицы

Вероятность → при ↓		Лента = неверно	Лента = верно
Чемодан в самолете	Истекшее время		
Неверно	2	100	0
Неверно	3	100	0
Неверно	4	100	0
Неверно	5	100	0
Неверно	6	100	0
Неверно	7	100	0
Неверно	8	100	0
Неверно	9	100	0
Неверно	10	100	0
Верно	0	100	0
Верно	1	90	10
Верно	2	80	20
Верно	3	70	30
Верно	4	60	40
Верно	5	50	50
Верно	6	40	60
Верно	7	30	70
Верно	8	20	80
Верно	9	10	90
Верно	10	0	100

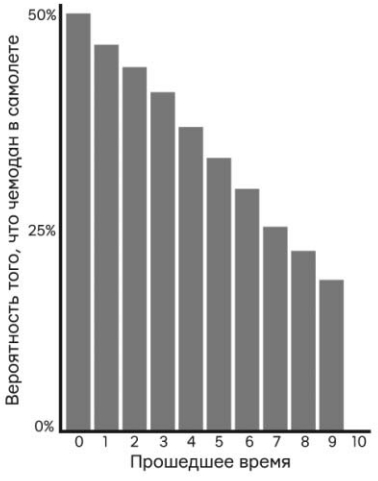


Рис. 17. Вероятность увидеть свой чемодан на ленте сначала снижается медленно, а потом быстрее (источник: график Маян Харел, информация Стефана Конради и Лайонела Джоффа)

Мы не только получили практический совет, но и поняли, что не стоит делать такие вещи в уме. Даже в крошечной сети с тремя узлами оказалось $2 \cdot 11 = 22$ родительских состояния, каждое из которых влияло на состояние потомка. Конечно, для компьютера эти вычисления элементарны, но… до определенного момента. Если делать их в организованной форме, сам объем вычислений может оказаться слишком большой нагрузкой даже для самого быстрого суперкомпьютера. Если у узла десять родителей, у каждого из которого два состояния, в таблице условной вероятности будет больше тысячи рядов. А если у каждого из 10 родителей 10 состояний, то в таблице будет 10 миллиардов рядов! По этой причине необходимо отсеять связи в сети, чтобы остались только самые важные и чтобы сеть была разреженной. Одним из технических достижений в развитии байесовских сетей стало выявление способов, которые позволяют использовать эту разреженность для сокращения времени вычислений.

Байесовские сети в реальной жизни

Сейчас байесовские сети — зрелая технология и готовое программное обеспечение для них можно купить у нескольких компаний. Байесовские сети также встроены во многие «умные» устройства. Чтобы дать вам представление о том, как они используются на практике, давайте вернемся к программе *B*onaparte для сравнения ДНК, с которой мы начали эту главу.

В Нидерландском институте судебной экспертизы эту программу используют каждый день, в основном расследуя дела о пропавших без вести, уголовные преступления и иммиграционные вопросы (желающие переехать в Нидерланды в статусе беженца должны доказать, что у них есть 15 родственников, живущих в стране). Однако байесовские сети продемонстрировали самый впечатляющий результат после катастрофы, такой как крушение рейса MH17 «Малайзия эйрлайнс».

Почти никого из жертв авиакатастрофы не удалось идентифицировать, сравнив ДНК с места катастрофы с ДНК из центральной базы данных. Следующим логичным шагом было взять у родственников образцы ДНК и искать частичные совпадения с ДНК жертв. Традиционные (небайесовские) методы позволяют это сделать, и они сыграли важнейшую роль в раскрытии нескольких давних преступлений в Нидерландах, США и других странах. Например, простая формула под названием «индекс отцовства» или «индекс сиблинга» помогает оценить вероятность того, что не идентифицированная ДНК принадлежит сыну или брату человека, чья ДНК есть у экспертов.

Однако эти индексы дают ограниченный результат, потому что они работают только для одного типа родства и только для близких родственников. Идея *B*onaparte состоит в том, чтобы можно было использовать данные о ДНК более дальних родственников или от нескольких родственников сразу. *B*onaparte делает это, преобразовывая родословную семьи в байесовскую сеть (рис. 18).

На рис. 19 мы видим, как *B*onaparte переводит один небольшой кусочек родословной в (причинную) байесовскую сеть. Главная проблема состоит в том, что генотип индивида, кото-

рый определяет генетическая экспертиза, содержит элементы, полученные и от отца, и от матери, но мы не можем определить их происхождение. Таким образом, два этих элемента (которые называются «аллели») необходимо рассматривать как скрытые, неизмеримые переменные в байесовской сети. Часть задачи состоит в том, чтобы вывести вероятность причины (ген голубых глаз был унаследован от отца) из имеющейся информации (например, есть гены голубых глаз и черных глаз; у кузенов со стороны отца голубые глаза, но у кузенов со стороны матери черные глаза).

Это задача на определение обратной вероятности, для чего и было изобретено правило Байеса.

Рис. 18. Фактическая родословная семьи с несколькими погибшими в авиакатастрофе рейса MH17 «Малайзия эйрлайнс» (источник: данные предоставлены Виллемом Бургерсом)

После того как байесовская сеть построена, финальный шаг — ввести ДНК жертвы и вычислить вероятность того, что она занимает определенное место в генеалогическом древе. Это делается путем распространения убеждений с помощью правила Байеса. Сеть начинается с определенной степени уверенности в каждом возможном утверждении об имеющихся в ней узлах, например: «отцовская аллель цвета глаз у этого человека — голубая». По мере того как в сеть вводится новая информация — неважно, в какое место, — степени уверенности в каждом узле, вверх и вниз по сети, будут меняться каскадно. Таким образом, как только мы обнаруживаем, что

данный образец является вероятным совпадением для одного человека в родословной, мы распространяем эту информацию вверх и вниз по сети. В итоге *B*onaparte учится не только на ДНК живых членов семьи, но и на уже полученных им результатах.

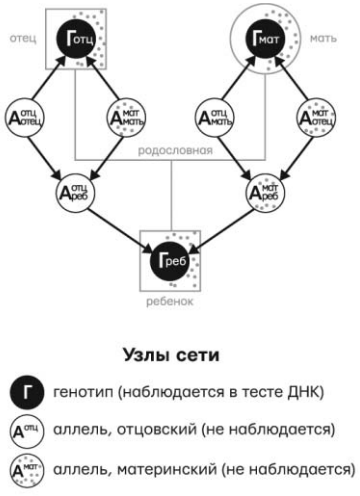


Рис. 19. От генетической экспертизы до байесовских сетей. В байесовской сети незакрашенные узлы представляют аллели, закрашенные — генотипы. Данные доступны только для закрашенных узлов, потому что генотипы не показывают, какая аллель была унаследована от отца, какая от матери. Байесовская сеть позволяет сделать выводы о ненаблюдаемых узлах, а также оценить вероятность того, что данный образец ДНК был получен от ребенка (источник: инфографика Маяна Харела)

Этот случай живо иллюстрирует преимущества байесовских сетей. Как только сеть настроена, следователю не нужно вмешиваться, чтобы сообщить ей, как оценить новые данные. Обновление происходит очень быстро (байесовские сети особенно хороши для программирования на распределенных вычислительных системах). Сеть интегративна, т.е. вся она реагирует на новую информацию. Вот почему даже ДНК тети или троюродного брата может помочь в идентификации жертвы. Байесовские сети — почти живая органическая ткань,

и неслучайно именно эту картину я держал в уме, пока пытался добиться, чтобы они заработали. Я хотел, чтобы байесовские сети работали как нейроны в человеческом мозге: когда касаешься одного нейрона, реагирует вся сеть, распространяя информацию на все остальные клетки в системе.

Прозрачность байесовских сетей отделяет их от большинства других подходов к машинному обучению, которые часто производят непроницаемые «черные ящики». В байесовской сети вы можете проследить каждый шаг и понять, почему те или иные данные изменили уверенность сети.

Какой бы изящной ни была программа *B*onaparte, она ничего не стоит без одной способности, которой не располагает (пока), — человеческой интуиции. Программа проводит анализ и сообщает специалистам, кому мог принадлежать каждый образец ДНК, составив рейтинг самых вероятных вариантов, а также сообщает о коэффициенте вероятности. После этого эксперты объединяют информацию о ДНК с данными о вещественных доказательствах, найденных на месте крушения, и делают окончательные выводы, не без помощи интуиции. Пока компьютер не может провести идентификацию самостоятельно. Одна из целей причинного вывода — создать более удобный интерфейс для взаимодействия человека и машины, который позволит включить интуицию следователя в процесс распространения убеждений.

Пример с генетической экспертизой дает самое поверхностное представление о том, как байесовские сети можно применять в геномике. Однако я хотел бы перейти к следующей области их применения, которая стала повсеместной в современном обществе. Более того, есть хорошие шансы, что у вас есть байесовская сеть в кармане прямо сейчас. Она называется «сотовый телефон». В каждом таком устройстве используются алгоритмы исправления ошибок, основанные на распространении степени уверенности.

Начнем с самого начала: когда вы говорите по телефону, он преобразует ваш прекрасный голос в последовательность нулей и единиц (которые называются биты) и трансформирует их, используя радиосигнал. К сожалению, ни один из них

не принимается со 100%-ной точностью. Пока он идет от башни сотовой связи и до телефона вашего друга, отдельные биты сменятся с нуля на единицу или наоборот.

Для исправления этих ошибок можно добавить избыточную информацию. Простейшая схема состоит в том, чтобы повторить каждый бит информации три раза: закодировать единицу как 111 и ноль как 000. Допустимые строки 111 и 000 называются кодовыми словами. Если приемник получит недопустимую строку, например 101, он будет искать наиболее вероятное допустимое кодовое слово, чтобы объяснить ее. Здесь, скорее всего, ошибка — ноль, а не две единицы, поэтому декодер интерпретирует это сообщение как 111 и, таким образом, заключит что бит был единицей.

К сожалению, это крайне неэффективное кодирование, потому что все наши сообщения становятся в три раза длиннее. Однако специалисты по телекоммуникациям 70 лет работают над кодами исправления ошибок, постоянно их улучшая.

Проблема декодирования аналогична другим проблемам с обратной вероятностью, которые мы обсудили, потому что мы снова хотим вывести вероятность гипотезы (послали сообщение Hello *W*orld!) из имеющихся данных (получили сообщение H*X*llo *W*ovld!). Кажется, пришло время применить распространение убеждений.

В 1993 году инженер «Франс телеком» по имени Клод Берру поразил мир программирования кодом для исправления ошибок, который позволял добиться почти идеальных результатов (другими словами, требуемый объем лишней информации был близок к теоретическому минимуму). Его идею под названием «турбокод» можно проиллюстрировать, представив ее с помощью байесовской сети.

На рис. 20а показано, как работает обычный код. Биты информации, которые поступают в телефон, когда вы говорите, показаны в первом ряду. Они кодируются любым способом — назовем его код *A* — в кодовые слова (второй ряд), которые потом принимаются с некоторыми ошибками. Это диаграмма — байесовская сеть, и мы можем использовать распространение

убеждений, чтобы вывести из полученных битов, каковы были биты информации. Однако это никак не улучшит код A.

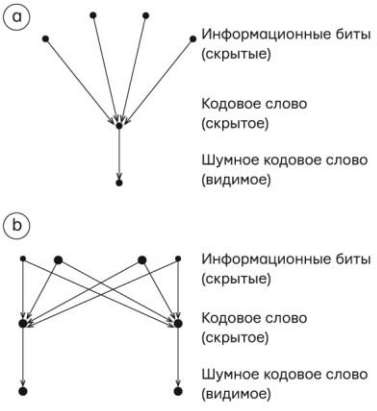


Рис. 20. Представление обычного процесса кодирования и турбокода в виде байесовской сети: *а* — информационные биты преобразуются в кодовые слова; они передаются и принимаются в пункте назначения с шумом (ошибками); *б* — информационные биты скремблируются и кодируются дважды. Декодирование происходит путем распространения убеждений в этой сети. Каждый процессор внизу использует информацию от другого процессора, чтобы улучшить предположение о скрытом кодовом слове в итеративном процессе

Блестящая идея Берру состояла в том, чтобы закодировать каждое сообщение дважды: один раз непосредственно и один раз уже после того, как сообщение скремблировано (преобразовано в случайную последовательность). Это приводит к созданию двух отдельных кодовых слов и получению двух шумных сообщений (рис. 20б). Нам неизвестна формула, которая позволяет напрямую декодировать такое двойное сообщение. Но Берру на опыте показал, что, если неоднократно примерить формулы распространения убеждений к байесовским сетям, происходят две потрясающие вещи. В большинстве случаев (и здесь я имею в виду около 99,999% случаев) вы получаете верные информационные биты. Более того, можно использо-

вать гораздо более короткие кодовые слова. Проще говоря, две копии кода А гораздо лучше одной.

Эта небольшая история верна, за исключением одного: Берру не знал, что работает с байесовскими сетями! Он просто сам открыл алгоритм распространения убеждений. И только пять лет спустя Дэвид Маккей из Кембриджа понял, что это тот же алгоритм, с которым он развлекался в конце 1980-х, рассматривая байесовские сети. Это поместило алгоритм Берру в знакомый теоретический контекст и позволило информатикам-теоретикам лучше понять его работу.

Вообще-то, другой инженер, Роберт Галлагер из Массачусетского технологического института, открыл код, в котором использовалось распространение убеждений (хотя его тогда не называли этим термином), еще в 1960 году, так давно, что Маккей описывает его код как «почти ясновидение». В любом случае он слишком опережал свое время. Галлагеру требовались тысячи процессоров на чипе, которые передавали туда и обратно сообщения о степени уверенности в том, что конкретный информационный бит равен единице или нулю. В 1960 году это было невозможно, и его код был практически забыт, пока Маккей не открыл его заново в 1998 году. Сегодня он есть в каждом сотовом телефоне.

Как бы то ни было, турбокоды имели ошеломляющий успех. До турбореволюции сотовые сети 2G использовали «мягкое декодирование» (т.е. вероятности), а не распространение убеждений. В сетях 3G применили турбокоды Берроу, а в 4G — турбокоды Галлагера. С точки зрения потребителя это означает, что ваш телефон потребляет меньше энергии, а аккумулятор работает дольше, потому что кодирование и декодирование — самые энергоемкие процессы. Кроме того, более совершенные коды означают, что не нужно находиться как можно ближе к вышке сотовой связи, чтобы получить высококачественную передачу. Другими словами, байесовские сети позволили производителям телефонов выполнить обещание: больше полосок в больше мест.

От байесовских сетей к диаграммам причинности

Возможно, после главы, посвященной байесовским сетям, у вас возник вопрос: как они относятся к остальному в этой книге, в частности к диаграммам причинности вроде тех, что приведены в главе 1? Конечно, я обсудил их так подробно отчасти потому, что они привели к причинности лично меня. Но, что еще важнее как с теоретической, так и с практической точки зрения, байесовские сети — ключ, который позволяет диаграммам причинности взаимодействовать с данными. Все вероятностные свойства байесовских сетей (включая связки, которые мы обсуждали выше в этой главе) и разработанные для них алгоритмы распространения убеждений подходят и для диаграмм причинности. Более того, они необходимы для понимания причинного вывода.

Основные различия между байесовскими сетями и диаграммами причинности заключаются в том, как они построены и в каких целях используются. Байесовская сеть — это всего лишь компактное представление огромной таблицы вероятностей. Стрелки означают, что вероятности дочерних узлов связаны со значениями родительских узлов определенной формулой (таблицы условных вероятностей) и что этого отношения достаточно, т.е. знание дополнительных родителей не изменит формулу. Точно так же отсутствие стрелки между любыми двумя узлами означает, что они независимы, если нам известны значения их родителей. Мы видели простую версию этого утверждения выше, когда обсуждали эффект экранирования в цепях и звеньях. В цепочке $A \rightarrow B \rightarrow C$ отсутствующая стрелка между A и C означает, что A и C независимы, если мы знаем значения их родителей. Поскольку у A нет родителей, а единственный родитель C — это B, отсюда следует, что A и C независимы, если мы знаем значение B, что согласуется со сказанным выше.

Однако если бы та же диаграмма была сделана как диаграмма причинности, то и замысел, который лежал бы в ее основе,

и пути интерпретации изменились бы. На этапе создания нужно рассмотреть каждую переменную, скажем *C*, и спросить себя, какие другие переменные она «слушает», прежде чем выбрать значение. Цепочка $A \rightarrow B \rightarrow C$ означает, что *B* слушает только *A, C* слушает только *B* и *A* не слушает; т.е. она определяется внешними силами, которые не входят в нашу модель.

Эта метафора слушания обобщает все знания, которые передает причинная сеть; остальные можно вывести, иногда для этого понадобятся данные. Обратите внимание, что, если мы изменим порядок стрелок в цепочке, таким образом получив $A \leftarrow B \leftarrow C$, причинное прочтение структуры резко изменится, но условия независимости останутся прежними. Отсутствие стрелки между *A* и *C* по-прежнему будет означать, что *A* и *C* независимы, если нам известно значение *B*, как в исходной цепочке. Из этого вытекают два чрезвычайно важных следствия. Во-первых, причинные допущения не изобретаются по нашей прихоти; они подвергаются тщательной проверке данными и могут быть сфальсифицированы. Например, если наблюдаемые данные не показывают, что *A* и *C* являются независимыми при наличии *B*, то мы вправе с уверенностью сделать вывод, что модель цепочки несовместима с данными и ее необходимо отбросить (или исправить). Во-вторых, графические свойства диаграммы определяют, какие модели причинно-следственных связей различают по данным, а какие навсегда останутся неразличимыми, независимо от объема данных. Так, мы не в состоянии отличить вилку $A \leftarrow B \rightarrow C$ от цепочки $A \leftarrow B \leftarrow C$ только по данным, потому что две диаграммы подразумевают одинаковые условия независимости.

Еще один удобный способ осмыслить каузальные модели — представить их в виде гипотетических экспериментов. Каждую стрелку можно считать утверждением об итоге гипотетического эксперимента. Стрелка от *A* к *C* означает, что если мы в силах повлиять только на *A*, то будем ожидать, что вероятность *C* изменится. Отсутствующая стрелка от *A* к *C* означает, что в том же эксперименте мы не увидим изменений в *C*, если сохраним родителей C неизменными (другими словами, *B* в примере выше). Обратите внимание на то, что в первом

случае мы рассуждали в терминах вероятности («если нам известно значение *B*»), а теперь в терминах причинно-следственных связей («если мы сохраним *B* неизменным»), а это подразумевает, что мы физически оградим *B* от изменений и отключим стрелку от *A* к *B*.

Причинные рассуждения, необходимые для создания каузальной сети, конечно же, дадут результат, расширив группу вопросов, на которые она может ответить. В то время как байесовская сеть способна всего лишь рассказать, насколько вероятно одно событие, если мы наблюдаем другое (информация первого уровня), диаграммы причинности в состоянии ответить на вопросы об интервенции и контрфактивные вопросы. Например, вилка $A \leftarrow B \rightarrow C$ однозначно сообщает нам, что, если «пошевелить» *A*, это не окажет никакого эффекта на *C*, каким бы интенсивным ни было шевеление. Однако байесовская сеть не рассчитана на учет шевелений и не позволяет увидеть разницу между наблюдением и действием или в самом деле отличить вилку от цепочки. Другими словами, и вилка, и цепочка показали бы, что наблюдаемые изменения в *A* ассоциируются с изменениями в *C*, не давая предсказаний об эффекте воздействия на *A*.

Теперь мы переходим ко второму, возможно, более важному эффекту байесовских сетей на причинный вывод. Открытые нами отношения между графической структурой диаграммы и данными, которые она представляет, теперь помогают нам моделировать шевеления, не делая этого физически. В частности, последовательно используя обусловливание, мы предскажем эффект действий или интервенций, не проводя собственно эксперимент. Чтобы это продемонстрировать, снова рассмотрим причинную вилку $A \leftarrow B \rightarrow C$, для которой мы сочли корреляцию *A* и *C* ложной. Это реально подтвердить экспериментом, в котором мы шевелим *A* и не находим корреляции *A* и *C*. Но можно все сделать лучше. Для этого нужно попросить диаграмму эмулировать эксперимент и сказать нам, способно ли ограничение по определенному параметру воспроизвести корреляцию, которая будет преобладать в эксперименте. Ответ последует

положительный: «Корреляция между *A* и *C*, измеренная после ограничения по B, окажется равной корреляции, которую мы увидим в эксперименте». Эту корреляцию можно оценить, использовав данные, и в приведенном случае она будет нулевой, что адекватно подтверждает наш интуитивный вывод: пошевелив *A*, мы не окажем никакого воздействия на *C*.

Эта способность эмулировать интервенции с помощью умных наблюдений не была бы достигнута, если бы не статистические свойства байесовских сетей, которые были обнаружены между 1980 и 1988 годами. Теперь мы решаем, какой набор переменных необходимо измерить, дабы предсказать эффект интервенций на базе наблюдений. Также мы в состоянии ответить на вопрос «Почему?». Например, кто-то спросит: почему воздействие на *A* заставляет *C* меняться? Действительно ли это прямой эффект *A* или это эффект медиации от переменной *B*? Если это и то и другое, можем ли мы оценить, какая доля этого эффекта обусловлена *B*?

Чтобы ответить на такие вопросы о медиации, надо предвидеть две одновременные интервенции: когда мы изменяем *A* и сохраняем *B* постоянным (чтобы отличить от обусловливания по *B*). Если нам удастся осуществить эту интервенцию физически, то мы получим ответ на наш вопрос. Но, будучи зависимыми от наблюдательных исследований, мы должны имитировать два эти действия с помощью ряда осознанных наблюдений. И вновь графическая структура диаграммы подскажет нам, возможно ли это.

Все это еще не было открыто в 1988 году, когда я начал размышлять, как объединить причинность с диаграммами. Я знал только, что байесовские сети в существовавшей тогда форме не могли ответить на вопросы, которые я задавал. Осознание того, что на основании одних лишь данных нельзя даже отличить $A \leftarrow B \rightarrow C$ от $A \rightarrow B \rightarrow C$, было источником боли и фрустрации.

Я знаю, что вам, читатель, уже не терпится узнать, как диаграммы причинности позволяют нам делать вычисления вроде тех, которые я только что описал. И мы туда доберемся — в главах с седьмой по девятую. Но пока мы не готовы, потому, начиная говорить о наблюдательных и экспериментальных

исследованиях, мы тут же покидаем мирные воды — сферу искусственного интеллекта — и сразу погружаемся в бурные воды статистики, которые вспенились после несчастливого расставания с причинностью. В ретроспективе оказалось, что борьба за принятие байесовских сетей в сфере ИИ была приятной прогулкой — да нет, роскошным круизом! — по сравнению с боем за диаграммы причинности, который мне пришлось вынести потом. И эта битва идет до сих пор — еще остались островки сопротивления.

Чтобы ориентироваться в этих новых водах, нужно будет понять способы, которыми традиционные статистики научились справляться с причинно-следственные связями, и ограничения этих способов. Поднятые выше вопросы о результатах интервенции, включая прямые и косвенные эффекты, не рассматриваются в традиционной статистике прежде всего потому, что отцы-основатели этой науки очистили ее от языка причин и следствий. Но статистики тем не менее считают допустимым говорить о причинах и следствиях в ситуации рандомизированного контролируемого исследования, в котором препарат *A* случайным образом назначается одним людям, а не другим, а затем сравниваются наблюдаемые изменения в *B*. Здесь и традиционная статистика, и наука о причинном выводе воспринимают предложение «*A* вызывает *B*» в одном и том же смысле.

Прежде чем мы обратимся к новой науке причин и следствий, проиллюстрированной каузальными моделями, сначала надо попытаться понять сильные стороны и ограничения старой, слепой к моделям науке. Чтобы прийти к выводу «*A* является причиной B», необходима рандомизация, и РКИ важны для нейтрализации «осложнителей» — вмешивающихся факторов (причины этого и природу обозначенной угрозы мы рассмотрим в следующей главе). По моему опыту, для большинста статистиков, а также для современных специалистов по анализу данных это не самые удобные вопросы, потому что их нельзя сформулировать, используя ориентированный на данные словарь. Более того, они часто не соглашаются по поводу того, что такое осложнение.

Исследовав эти вопросы в свете диаграмм причинности, мы можем поместить РКИ в соответствующий контекст. Целесообразно или рассматривать их как особый случай причинного вывода, или считать причинный вывод сильным расширением РКИ. Любой подход из этих двух имеет право на существование, и не исключено, что люди, которых приучили видеть в РКИ способ определить причинность, найдут второй более подходящим.

Глава 4
Осложнители и наоборот: как убить прячущуюся переменную

Если бы наша концепция каузальных факторов хоть как-то была связана с рандомизированными исследованиями, последние были бы изобретены за пятьсот лет до Фишера.
Автор, 2016

Однажды у Асфеназа, придворного при царе Навуходоносоре, возникла большая проблема. В 597 году до н.э. вавилонский царь разорил Иудейское царство и привел с собой тысячи пленных, среди них и иерусалимскую знать. По обычаю своего царства, он возжелал, чтобы некоторые из них служили ему при дворе, и он приказал Асфеназу сыскать среди них «отроков, у которых нет никакого телесного недостатка, красивых видом и понятливых для всякой науки, и разумеющих науки и смышленых и годных служить в чертогах царских». Этим счастливцам предстояло изучить язык и культуру Вавилона и служить в администрации великой империи, раскинувшейся от Персидского залива до Средиземного моря. На время обучения царь назначил им пищу с царского стола и вино, которое сам пил.

Тут-то и возникла проблема. Один из любимцев вельможи, юноша по имени Даниил, отказался притрагиваться к пище. По религиозным мотивам он не мог есть мясо, приготовлен-

ное не по иудейским обычаям, и он попросил, чтобы ему и его товарищам взамен позволили питаться растительной пищей.

Асфеназ рад был бы удовлетворить его просьбу, но боялся, что это заметит царь: «… Если он увидит лица ваши худощавее, нежели у отроков, сверстников ваших, то вы сделаете голову мою виновною пред царем».

Даниил постарался убедить Асфеназа, что диета из воды и овощей не уменьшит их способность служить царю. Как и подобает «разумеющему науки и смышленому», он предложил эксперимент: «Сделай опыт над рабами твоими, — предложил он, — в течение десяти дней пусть дают нам в пищу овощи и воду для питья. И потом пусть явятся пред тобою лица наши и лица тех отроков, которые питаются царской пищею…» И сказал Даниил: «… Поступай с рабами твоими, как увидишь».

Даже если вы не читали эту историю, то, вероятно, уже догадались, что случилось. Даниил и трое его друзей прекрасно чувствовали себя на вегетарианской диете. Царь был настолько поражен их умом и способностью к ученью — не говоря уж об их здоровом, цветущем виде, — что дал им лучшее место при дворе, где «во всяком деле мудрого уразумения… находил их в десять раз выше всех тайноведцев и волхвов, какие были во всем царстве его». Позже Даниил прославился толкованием снов царя и выжил, брошенный в ров со львами.

Библейской истории можно верить либо не верить, но она прекрасно передает суть сегодняшней экспериментальной науки. Асфеназ задает вопрос о причинности: отощают ли мои слуги на вегетарианской диете? Даниил предлагает методологию для работы со всеми подобными вопросами: возьмите две группы людей, одинаковые по всем важным для нас параметрам. Поместите одну из них в новые условия (задайте особую диету, давайте лекарство и т.п.), а вторую группу (называемую контрольной) оставьте в старых условиях (не давайте лекарства и т.п.). Если после подходящего отрезка времени между этими предположительно одинаковыми группами людей наблюдается измеримая разница, тогда новые условия должны быть ее причиной.

Теперь мы называем это контролируемым исследованием. Принцип прост. Чтобы понять каузальный эффект диеты, в идеале нам надо бы сравнить то, что произойдет с Даниилом на одной диете, с тем, что произойдет с ним же на другой, но мы не можем вернуться в прошлое и переписать историю, поэтому делаем лучшее из того, что нам остается: мы сравниваем подопытную группу с контрольной. Очевидно, но при этом очень важно, что группы должны быть сравнимы между собой, им необходимо репрезентативно представлять какую-либо популяцию. Если эти условия соблюдены, результаты могут быть перенесены на популяцию в целом. К чести Даниила, он, похоже, это понимал. Он не просит овощей только для себя: если опыт покажет, что вегетарианская диета лучше, тогда эта диета в будущем будет дозволена всем невольникам-израильтянам. По крайней мере, так я интерпретирую фразу «Поступай с рабами твоими, как увидишь».

Даниил понимал также и то, что сравнения нужно проводить между группами. В этом смысле он уже был мудрее многих наших современников, которые выбирают, например, модную диету только потому, что какая-то их подруга села на нее и похудела. Если вы выбираете способ питания, основываясь только на опыте одной подруги, вы предполагаете тем самым, что являетесь одинаковой с ней по всем важным в данном случае показателям: росту, весу, наследственности, условиям жизни, предшествовавшим диетам и т.д. Это очень большое допущение.

Другой ключевой момент в эксперименте Даниила в том, что он предполагался в будущем, а экспериментальные группы были подобраны заранее. По контрасту представьте, что вы смотрите рекламный ролик, в котором 20 человек все как один рассказывают, что сбросили вес на такой-то диете. Со стороны это довольно приличная выборка, поэтому некоторым зрителям ролик кажется убедительным. Но на самом деле это означает, что принимать решение придется, основываясь только на опыте людей, для которых она заведомо сработала. Из жизненного опыта вы знаете, что на одного человека, которому диета помогла, приходится десяток таких же, как он или она, которые тоже ее пробовали, но безрезультатно. Но их, конечно, не станут показывать в рекламном ролике.

По всем этим показателям эксперимент Даниила был очень современным. Контролируемые заранее планируемые эксперименты по-прежнему остаются фирменным знаком хорошей науки. Однако об одной важной вещи Даниил не подумал: это систематическая ошибка. Допустим, что Даниил и его друзья изначально здоровее, чем контрольная группа. В этом случае их здоровый внешний вид на десятый день диеты может быть никак не связан с самой диетой: он лишь отражает их общее здоровье. Не исключено, что они выглядели бы даже лучше, если бы согласились есть мясо со стола царя!

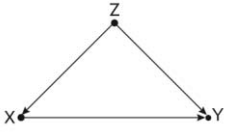


Рис. 21: Самая базовая версия путаницы: Z — это путаница в предполагаемой причинно-следственной связи между X и Y.

Систематическая ошибка наблюдается, когда некая переменная влияет на представителей одновременно и подопытной, и контрольной групп. Иногда эти вмешивающиеся переменные известны; в других случаях о них можно только догадываться и они действуют как скрытые переменные. На каузальной диаграмме вмешивающиеся переменные, или конфаундеры, легко распознать: на рис. 21 переменная Z в центре вилки осложняет переменные X и Y (позже мы увидим более универсальное определение, но такой треугольник — самая узнаваемая и распространенная ситуация). Из диаграммы легко понять, почему вмешивающуюся переменную называют конфаундером. Истинный каузальный эффект $X \to Y$ осложняется ложной корреляцией между ними, возникающей в результате вилки $X \leftarrow Z \to Y$. Например, если мы тестируем некое лекарство и даем его пациентам, которые в среднем моложе, чем контрольная группа, возраст становится конфаундером — скрытой третьей переменной. Если у нас нет данных по возрастам ис-

пытуемых, мы не сможем отделить истинный эффект нашего препарата от ложного эффекта.

Однако верно и обратное. Если у нас есть измерения третьей переменной, то разделить истинный и ложный эффекты становится очень просто. Так, если вмешивающаяся переменная — это возраст, мы сравниваем опытную и контрольную группу в каждой возрастной группе отдельно. Затем можно усреднить воздействие, подсчитывая вес каждой группы соответственно ее проценту в целевой популяции. Этот метод компенсации знаком всем статистикам: он называется «корректировка по *Z*» или «поправка по *Z*».

Как ни странно, статистики одновременно и недо-, и переоценивают важность корректировки по конфаундеру. Переоценка заключается в том, что поправки вводятся по слишком многим переменным или даже по переменным, по которым их вводить неправильно. Недавно я наткнулся на цитату из политического блогера по имени Эзра Кляйн, в которой эта гиперкорректировка описана весьма точно: «В статьях это попадается постоянно. „Мы скорректировали данные по…". Далее следует список, чем он длиннее, тем лучше. Уровень дохода. Возраст. Раса. Религия. Рост. Цвет волос. Сексуальные предпочтения. Регулярность посещения спортзала. Любовь к родителям. Кока-кола или пепси-кола. Чем больше корректировок, тем значительнее ваша статья. Ну или, по крайней мере, тем значительней она выглядит. Поправки дают ощущение конкретики, точности. Но иногда поправок слишком много, и в результате вы корректируете как раз то, что хотите измерить». Кляйн затрагивает важную тему. В статистике давно образовалось непонимание того, какие переменные следует, а какие не следует корректировать, поэтому по умолчанию поправки стали вводить для всего, что только можно измерить. Подавляющее большинство современных работ поддерживают эту практику. Это удобная и несложная процедура, но она одновременно тратит впустую время и создает ошибки. Ключевое достижение Революции Причинности в том, что она положила конец этой путанице.

В то же время статистики сильно недооценивают корректировку в том смысле, что вообще избегают говорить о при-

чинности, даже если все поправки сделаны верно. Это тоже противоречит основной идее этой главы: если вы обнаружили значительный набор вмешивающихся переменных в диаграмме, получили по ним данные и ввели по ним поправки, то у вас есть полное право сказать, что вы подсчитали причинностное воздействие $X \to Y$ (при условии, конечно, что ваша диаграмма научно обоснована).

Подход учебников по статистике к вмешивающимся переменным совершенно иной, он опирается на идею, наиболее активно защищаемую Р. Э. Фишером: рандомизированное контролируемое исследование. Ученый был совершенно прав в этом подходе, но не в его основаниях. РКИ — это действительно замечательное изобретение, но до недавнего времени поколения статистиков, следуя за Фишером, неспособны были доказать, что то, что они получали благодаря РКИ, было именно тем, что они хотели получить. У них не было языка, с помощью которого можно было бы записать то, что они хотели найти, а именно каузальное воздействие X на Y. В этой главе одна из моих целей — объяснить с точки зрения каузальных диаграмм, почему именно РКИ позволяют нам оценить каузальное воздействие $X \to Y$, не становясь жертвой систематической ошибки. Когда мы поймем, как именно работают РКИ, нам не нужно будет больше помещать их на пьедестал и относиться к ним как к золотому стандарту причинностного анализа, который все остальные методы должны воспроизводить. Совсем наоборот: мы увидим, что так называемый золотой стандарт легитимен потому, что опирается на более базовые принципы.

Эта глава также покажет, что каузальные диаграммы позволяют переключаться с конфаундеров на деконфаундеры (*deconfounders*). Первые вызывают проблемы — вторые решают ее. Они могут перекрываться, но это не обязательно. Если у нас есть данные по достаточному набору деконфаундеров, не будет иметь значения, если мы проигнорируем некоторые или даже все конфаундеры.

Это переключение внимания — основной путь, по которому Революция Причинности позволяет нам продвинуться дальше фишеровских экспериментов и выявить причинно-следствен-

ные связи из неэкспериментальных исследований. С помощью него реально определить, какие переменные должны быть скомпенсированы, чтобы стать деконфаундерами. Этот вопрос десятилетиями терзал как теоретиков, так и практиков статистики; десятилетиями здесь скрывалась ахиллесова пята всей отрасли знания. Так происходило потому, что он не имеет никакого отношения ни к данным, ни к статистическим методам. Конфаундеры — это причинностная концепция, она находится на второй ступени Лестницы Причинности.

Графические методы, возникнув в 90-е годы прошлого века, полностью упростили проблему конфаундеров. В частности, скоро мы познакомимся с методом критерия черного хода, который недвусмысленно определяет, какие переменные в каузальной диаграмме являются деконфаундерами. Если исследователь в состоянии получить данные по этим переменным, он может скорректировать их влияние и предсказать результаты действия, даже не осуществляя его.

На самом деле Революция Причинности идет дальше. В некоторых случаях мы вправе ввести поправку по конфаундерам даже тогда, когда у нас нет данных по достаточному массиву деконфаундеров. В этих случаях целесообразно использовать другие формулы корректировки — не общепринятые, которые работают только с критерием черного хода — и убрать всю систематическую ошибку. Эти впечатляющие разработки мы прибережем для главы 7.

Хотя вмешивающиеся переменные известны очень давно во всех областях науки, понимание, что эта проблема требует каузальных, а не статистических решений, пришло относительно недавно. Даже совсем недавно, в 2001 году, рецензенты отклонили мою статью, настаивая на том, что «проблема вмешивающихся переменных целиком лежит в плоскости традиционной статистики». К счастью, за последнее десятилетие число таких редакторов резко сократилось. Теперь образовался практически всеобщий консенсус, по крайней мере среди философов, эпидемиологов и представителей общественных наук, в том, что: 1) проблема конфаундеров нуждается в каузальном решении и такое решение есть; 2) каузальные диаграммы —

это полный и систематический метод для нахождения таких решений. Эпоха сложностей с конфаундерами подошла к концу!

Леденящий ужас конфаундеров

В 1998 году в статье, опубликованной в «Медицинском журнале Новой Англии», сообщалось, что обнаружена связь между регулярными занятиями спортивной ходьбой и снижением смертности среди мужчин-пенсионеров. Исследователи использовали данные программы «Здоровье сердца» в Гонолулу, которая наблюдала за здоровьем 8 тысяч мужчин японского происхождения с 1965 года.

Исследователи во главе с Робертом Эбботом, специалистом по биологической статистике, хотели выяснить, живут ли дольше мужчины, занимающиеся физкультурой более регулярно. Они взяли выборку в 707 человек из более крупной группы в 8 тысяч, в которой все были достаточно здоровы физически для пеших прогулок. Группа Эббота выяснила, что за 12-летний период уровень смертности был в два раза выше среди мужчин, которые в день проходили менее мили (далее «малоходящие») по сравнению с теми, кто проходил больше 2 миль в день («многоходящие»). Точнее, среди малоходящих умерло 43,0%, в то время как среди многоходящих — только 21,5%.

Однако, поскольку экспериментаторы не выбирали случайным образом, кому из испытуемых предписывается ходить много, а кому мало, нам приходится учитывать возможность искажений из-за вмешивающихся факторов. Наиболее очевидный из них — возраст: более молодые пенсионеры, вероятно, более склонны к физическим нагрузкам, и одновременно вероятность смерти для них меньше. Таким образом, у нас получается каузальная диаграмма вроде той, что изображена на рис. 22.

Классическая вилка в узле «Возраст», говорит нам о том, что возраст — вмешивающаяся переменная для ходьбы и смертности. Я уверен, что вы в состоянии придумать и другие конфаундеры. Вероятно, малоходящие менее подвижны

не случайно, им просто трудно много ходить. Следовательно, состояние здоровья тоже может быть конфаундером. Подобный поиск вмешивающихся факторов способен продолжаться до бесконечности. Не исключено, что малоходящие больше пьют? Или чаще переедают?

Хорошие новости: исследователи постарались учесть все эти моменты. Они подсчитали и внесли поправки на такие факторы, как возраст, состояние здоровья, потребление алкоголя, особенности диеты и многие другие. Так, действительно оказалось, что многоходящие в среднем чуть моложе. Исследователи внесли поправки по возрасту и обнаружили, что разница в смертности между много- и малоходящими все еще остается значительной (41% для малоходящих при 24% для многоходящих).

Несмотря на это, исследователи все же были крайне осторожны в своих выводах. В конце статьи они писали: «Конечно, прямое влияние сознательных попыток увеличить расстояние, проходимое за день, на долголетие физически способных к ходьбе мужчин невозможно вывести из данных нашей работы». Используя язык главы 1, они отказываются что-либо говорить о вероятности выживания в следующие 12 лет при условии, что вы занимаетесь ходьбой — *do* (*ходьба*).

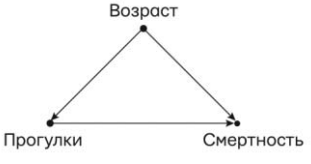


Рис. 22. Каузальная диаграмма для примера с ходьбой

Отдавая должное Эбботу и его группе, надо сказать, что у них действительно были причины для такой осторожности. Это было первое исследование на указанную тему, и выборка была относительно невелика и однородна. Тем не менее такая осторожность отражает более общую точку зрения, выходящую за пределы гомогенности и размеров выборки. Исследователей

приучили полагать, что работы, основанные на наблюдениях (такие, где испытуемые сами выбирают экспериментальные воздействия), не могут выявить действие каузальных факторов. Я считаю, что эта осторожность избыточна. Зачем еще прилагать усилия и вводить поправки по всем конфаундерам, если не для того, чтобы избавиться от ложной части связи и таким образом лучше понять каузальную часть? Вместо того чтобы говорить: «Конечно, мы не можем», как поступили они, нам следует провозгласить, что, разумеется, мы в состоянии кое-что сказать о намеренной интервенции. Если мы верим, что команда Эббота идентифицировала все важные конфаундеры, мы должны верить и тому, что намеренные занятия ходьбой непременно продлевают жизнь (по крайней мере, в случае японских мужчин).

Это прогностическое умозаключение, основанное на предположении, что никакие другие конфаундеры не играют сколько-нибудь значительной роли в выявленных отношениях переменных, — очень важная информация. Она точно сообщает потенциальному спортсмену, какого рода неопределенность остается, если принять это утверждение по номиналу. Она говорит, что остаточная неопределенность не выше, чем вероятность, что существуют дополнительные осложнители, которые не были приняты во внимание. Она также ценна тем, что определяет направление будущих исследований, которые должны сосредоточиться на этих других факторах (если они существуют), а не на тех, которые были нейтрализованы в данной работе. Короче говоря, знать набор допущений, которые стоят за данным выводом, не менее важно, чем пытаться обойти эти допущения при помощи РКИ, с которым как мы сейчас увидим, много своих сложностей.

Искусное дознание природы: почему РКИ работают

Как я уже говорил выше, есть одно обстоятельство, при котором ученые перестают избегать говорить о причинности: это происходит тогда, когда им удается провести рандомизиро-

ванное контролируемое исследование. Вы можете прочитать об этом в Википедии или в тысяче других мест: «РКИ часто считается золотым стандартом клинических испытаний». За это нам надо благодарить Р. Э. Фишера, так что весьма любопытно, что человек, очень близкий к нему, пишет о том, какие умозаключения привели его к этому. Цитата большая, но ее стоит привести полностью:

«Искусство и практика научного эксперимента целиком состоят в искусном допросе Природы. Наблюдение снабдило ученого видением Природы в некотором ее аспекте, у которого есть все недостатки добровольного показания. Он желает проверить верность своей интерпретации этого показания, для чего задает вопросы, нацеленные на установление причинно-следственных отношений. Его вопросы, в форме экспериментальных действий, в необходимой степени детальны, и он должен полагаться на последовательность Природы, делая общие выводы из ее ответа в отдельном случае или предсказывая исход на основании подобных операций в других случаях. Его цель — вывести обоснованные заключения определенной точности и уровня обобщения из полученных им показаний.

Природа, однако, ведет себя далеко не последовательно, ее ответы переменчивы, жеманны, двусмысленны. Она отвечает на вопрос в той форме, в которой он поставлен перед ней в эксперименте, а не в той, которая в голове у экспериментатора; она не собирается переводить ответы на понятный ему язык; ничем не делится даром; и она помешана на точности. Поэтому экспериментатор, который хочет, например, сравнить два удобрения, потратит время впустую, если, разделив свое поле на две равные части, удобрит одну одним, а вторую другим, затем засеет и сравнит собранный урожай между двумя половинами. Вопрос его задан так: какова разница между урожаем с участка А при условиях 1 и урожаем с участка Б при условиях 2? Он не спросил сначала, будет ли участок А давать урожай, одинаковый с участком Б при одинаковых условиях, и он не сможет отделить влияние свойств участка от влияния экспериментальных условий, поскольку Природа, в соответствии с запросом, записала не только вклады каждого из двух

различных удобрений в урожай, но и вклады, определяемые различиями между участками в плодородии почв, структуре, водоотведении, расположении, микрофлоре и сотнями других переменных».

Автор этого отрывка — Джоан Фишер Бокс, дочь Рональда Фишера, он взят из написанной ею биографии ее прославленного отца. Хотя сама она не посвятила себя статистике, она явно очень глубоко понимает главный вызов, с которым статистики сталкиваются. Она недвусмысленно утверждает, что вопросы, которые они задают, «нацелены на установление причинно-следственных связей». А то, что стоит у них поперек дороги, — это конфаундеры, хотя она и не употребляет этот термин. Они хотят узнать влияние удобрения (тогда говорили «унавоживания»), т.е. ожидаемую урожайность при применении одного удобрения в сравнении с урожайностью при применении альтернативы. Природа, однако, говорит им о влиянии удобрения в смеси (помните термин «вмешивающаяся переменная»?) со следствиями множества других причин.

Мне нравится образ, который Фишер Бокс предложила в процитированном отрывке: природа словно джинн из сказки, который отвечает точно на тот вопрос, который мы ему реально задали, а не на тот, который хотели бы задать. Но нам приходится верить (а Фишер Бокс, очевидно, верит), что ответ на тот вопрос, который мы хотим задать, действительно существует в природе. Наши эксперименты — это довольно неряшливый способ получить этот ответ, но они ни в коем случае не определяют его. Если мы точно следуем ее аналогии, то $do\ (X = x)$ должно быть сначала, потому что это свойство природы, представляющее искомый ответ: как повлияет на урожай применение первого удобрения на всем поле? Только затем идет рандомизация, потому что это присущий человеку способ получить ответ на данный вопрос. Можно сравнить ее с датчиком термометра, который представляет собой способ измерения температуры, но не саму температуру.

В молодые годы, работая на опытной станции в Ротамстеде, Фишер обычно применял очень сложный, систематический подход, для того чтобы отделить влияние удобрения от дру-

гих переменных. Он делил свои поля на сетку из небольших участков и тщательно планировал исследование так, чтобы каждое удобрение было испробовано с каждым опытным видом растений и типом почвы (рис. 23). Он проделывал это с целью получить уникальные образцы для сравнения их между собой; в реальности он никогда не смог бы предугадать все конфаундеры, способные определять плодородие данного участка. Достаточно умный джинн сможет победить любую самую совершенную схему структурирования поля.

Примерно в 1923 или 1924 году Фишер догадался, что единственный дизайн исследования, неподвластный «джинну», — это случайность. Представим, что мы ставим этот же самый эксперимент 100 раз на поле с неизвестным распределением плодородия почвы. Каждый раз вы назначаете то или иное удобрение для того или иного участка поля случайным образом. Иногда вам очень не везет, и вы назначаете удобрение 1 как раз на те участки, которые сами по себе наименее плодородны. В другой раз, наоборот, оно случайно попадает на плодородные участки. Но если вы свободно и случайно тасуете части поля при всякой следующей итерации эксперимента, можно гарантировать, что эффект везения или невезения нивелируется. В этом случае удобрение 1 будет назначено на определенной общей выборке участков поля, репрезентативно представляющей поле в целом. Это как раз то, что и нужно для контролируемого опыта. Поскольку распределение плодородности по полю остается одним и тем же во всех итерациях эксперимента — «джинн» не может его изменить, — он оказывается вынужден ответить (ну чаще всего!) на тот каузальный вопрос, который вы ему задали.

С нашей современной точки зрения в эпоху, когда рандомизированные опыты — это золотой стандарт, все вышесказанное может казаться очевидным. Но в то время сама идея случайности в схеме эксперимента привела коллег Фишера на статистическом поприще в откровенный ужас. Неприязнь усиливало, вероятно, и то, что Фишер буквально вынимал карты из тасованной колоды, назначая то или иное удобрение

для определенных участков поля. Подчинить науку причудам шанса — каково!

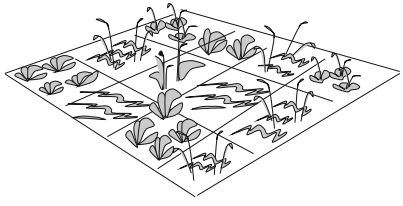


Рис. 23. Одна из множества придуманных инноваций Р. Э. Фишера — схема исследования «латинский квадрат», согласно которой один участок, засаженный данным типом растений, появляется в каждой строке (тип удобрения) и в каждом столбце (тип почвы). Подобные схемы все еще используются на практике, но Фишер затем убедительно показал, что рандомизированная схема еще более эффективна.

Однако Фишер хорошо понимал, что не очень точный ответ на правильный вопрос гораздо лучше, чем очень точный ответ на неверный вопрос. Если задавать «джинну» неправильные вопросы, вам никогда не выяснить у него то, что вы хотите знать. Если же вы ставите вопрос правильно, отдельные неверные ответы — гораздо меньшая проблема. Вы можете оценить, насколько эти ответы неточны, потому что неточность образуется в результате процедуры рандомизации (которая известна и понятна), а вовсе не из-за характеристик почвы на участках (которые неизвестны).

Таким образом рандомизация дает нам два преимущества. Первое — она элиминирует системную ошибку (благодаря ей мы правильно задаем вопрос природе). Во-вторых, она позволяет исследователю оценить неточность ответа. Тем не менее, согласно историку Стивену Стиглеру, Фишер ратовал за рандомизацию преимущественно из-за второго момента. В подсчете неточности, или, статистическим языком, ошибки, ему не было равных в мире, он разработал для этого множество новых математических процедур. При этом его понимание вмешивающихся переменных и их устранения было чисто

интуитивным, поскольку ему недоставало математической символики, для того чтобы адекватно передать то, что он искал.

Теперь, через 90 лет, мы можем воспользоваться оператором *do*, чтобы ответить на вопросы, которые Фишер хотел, но не мог задать. Давайте взглянем с каузальной точки зрения, каким образом рандомизация позволяет нам задать «джинну» правильный вопрос.

Начнем, как обычно, с каузальной диаграммы. Модель 1, показанная на рис. 24, показывает, как урожайность каждого участка определяется при нормальных условиях, когда фермер решает, как удобрять тот или иной участок, руководствуясь предвзятостью или прихотью. Вопрос, который он хочет задать джинну по имени Природа, таков: «Какова будет урожайность при однородном применении удобрения 1 (в сравнении с удобрением 2) на всем поле?». Или в терминах оператора *do*: каково P (*урожай do* (*удобрение* = 1))?

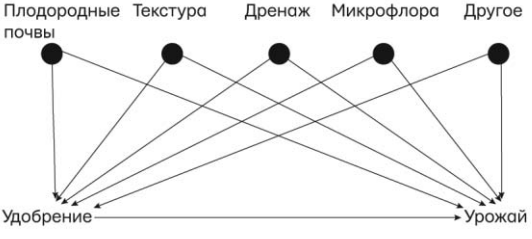


Рис. 24. Модель 1: неправильно контролируемое исследование

Если фермер ставит эксперимент наивно, например применяя удобрение 1 на верхней части поля, а удобрение 2 на нижней, то в качестве вмешивающейся переменной у него, вероятно, окажется дренированность. Если в один год он применит удобрение 1, а на другой — удобрение 2, то вмешивающейся переменной окажется погода. В любом случае сравнение окажется необъективным.

То, что хотел бы знать фермер, описывается моделью 2, когда все участки получают одно и то же удобрение (рис. 25). Как объяснялось в главе 1, действие оператора *do* — стереть все

стрелки, идущие к «удобрению», и придать этой переменной определенное значение, положим *удобрение* = 1.

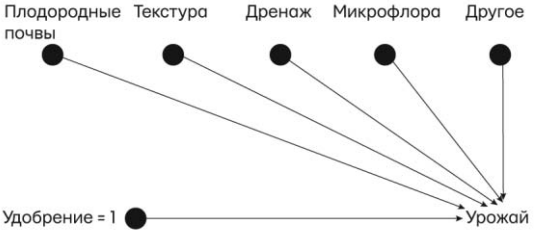


Рис. 25. То, что мы хотели бы знать

Наконец, давайте посмотрим, как все будет выглядеть после применения рандомизации. Теперь на некоторых участках поля будет *do* (*удобрение* = 1), а на других *do* (*удобрение* = 2), но выбор — какое воздействие будет оказано и на какой участок — окажется случайным. Эта ситуация описывается моделью 3 на рис. 26, в которой значение переменной *удобрение* назначается рандомизирующим устройством, например колодой карт, как у Фишера.

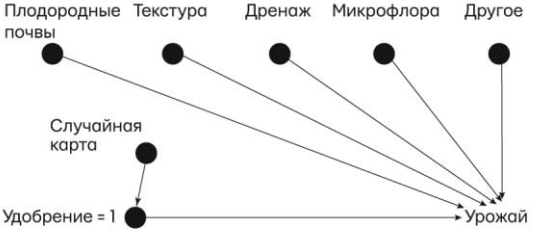


Рис. 26. Модель 3: ситуация, симулированная рандомизированным контролируемым опытом

Обратите внимание, что все стрелки, направленные к переменной *удобрение*, теперь исчезли, отражая предположение, что фермер в своем выборе удобрения на участке руководствуется только выпавшими картами. Не менее важно и то, что

от переменной *карта* к переменной урожайность нет стрелки, потому что растения не знают, что на карте (в случае растений это надежное допущение, но если рандомизированный опыт ставится на людях, об этом стоит подумать). Таким образом, модель 3 описывает ситуацию, в которой отношения между переменными *удобрение* и *урожайность* не осложнены (т.е. у них нет никакой общей причины). Это значит, что в ситуации на рис. 26 наблюдение *удобрение* = 1 — это то же самое, что и интервенция *удобрение* = 1.

Это подводит нас к ключевому выводу: рандомизация — это способ симулировать модель 2. Она убирает все имевшиеся конфаундеры, не внося новых. В этом источник ее силы: в ней нет ничего таинственного или мистического. Это всего лишь, как выразилась Джоан Фишер Бокс, «искусный допрос Природы». Эксперимент, однако, утратил бы свою объективность, если бы экспериментатор назначал бы удобрения по своему выбору или если бы растения на участке «знали», какая карта им выпала. Вот почему клинические исследования с участием людей приходится организовывать с большим вниманием к тому, чтобы информация о выборе была сокрыта от глаз как испытуемых, так и экспериментаторов (эта процедура называется двойным слепым исследованием).

Я добавлю к этому второй итоговый вывод: есть и другие способы симулировать ситуацию модели 2. Один из них состоит в том, чтобы выявить все конфаундеры, измерить их и внести по ним поправки. Однако рандомизация обладает одним огромным преимуществом: она обрывает все входящие каузальные связи к исследуемой переменной, включая те, о которых мы не знаем, и те, которые не в состоянии измерить (см. факторы «Другое» на рис. 24—26).

Напротив, в нерандомизированном эксперименте исследователь должен полагаться на свое знание предмета. Если он уверен, что его каузальная модель учитывает достаточное число факторов, устраняющих вмешивающиеся переменные, и по ним собрано достаточно данных, тогда возможно оценить влияние *удобрения* на *урожайность* объективно. Однако

всегда сохраняется опасность, что какая-то вмешивающаяся переменная осталась неучтенной и оценка окажется неточной.

При прочих равных рандомизированные контролируемые испытания по-прежнему предпочтительнее для научных работ, предполагающих наблюдение, точно так же, как сети безопасности рекомендуются для канатоходцев. Однако не все аналогии идеальны. В некоторых случаях интервенция неосуществима физически (например, в исследованиях о влиянии тучности на сердечные заболевания не получится рандомно назначить, кому из испытуемых быть тучным, а кому нет). В других интервенция неэтична (изучая влияние курения, мы не имеем права случайным образом назначить испытуемых, которые будут курить 10 лет). Кроме того, у нас могут возникнуть сложности с набором испытуемых для тестирования заведомо неприятных процедур, и добровольцы, которые в результате рискнут участвовать в эксперименте, не будут представлять собой репрезентативную выборку всей популяции в целом. К счастью, оператор *do* позволяет нам адекватным с точки зрения науки образом выявлять каузальные воздействия в неэкспериментальных научных работах, что бросает вызов традиционному главенству РКИ. Как обсуждалось выше в примере с ходьбой, подобные каузальные оценки, полученные из обзорных работ, целесообразно назвать условной причинностью, т.е. причинностью, условно возможной при наборе предположений, отраженных в нашей каузальной диаграмме. Важно, что мы не относимся к таким исследованиям как к гражданам второго сорта; у них есть то преимущество, что они проводятся в естественных условиях, на целевой аудитории, а не в искусственной среде лаборатории и они по-своему «чисты» — у них нет никаких проблем ни с этикой, ни с осуществимостью.

Теперь, когда мы понимаем, что принципиальный смысл РКИ в том, чтобы избавиться от осложнителей, давайте посмотрим и на другие методы, данные нам Революцией Причинности. История начинается со статьи 1984 года, написанной двумя моими давними коллегами, в которой было положено начало переоценке представлений о том, что такое конфаундеры.

Новая парадигма конфаундеров

«Хотя проблема конфаундеров, или вмешивающихся переменных, общепризнанно считается одной из центральных в эпидемиологических исследованиях, обзор литературы обнаруживает заметную непоследовательность в определении этих терминов», — этой фразой Сандер Гренланд из Калифорнийского университета в Лос-Анджелесе и Джейми Робинс из Гарвардского университета выразили самую суть причины, по которой в борьбе с конфаундерами ученые их времени не продвинулись ни на шаг со времен Фишера. Без сущностного понимания проблемы авторы обзорных работ, в которых контроль над условиями наблюдаемого невозможен, не могли сказать ничего осмысленного.

Каково было определение конфаундеров тогда и каким оно должно быть теперь? Благодаря современным знаниям о логике причинности, на второй вопрос ответить проще. То, что мы наблюдаем и способны измерить, — это вероятность данного исхода при данном воздействии, $P(Y \mid X)$. Вопрос, который мы задаем природе, имеет отношение к причинно-следственной связи между X и Y, которая выражается в интервенционной вероятности $P(Y \mid do(X))$. Конфаундеры, таким образом, должны быть определены просто как все, что приводит к несовпадению этих вероятностей: $P(Y \mid X) \neq P(Y \mid do(X))$. Что тут сложного?

К сожалению, до 90-х годов *XX* века все было непросто, потому что оператор *do* еще не был формализован. Даже сегодня, если вы поймаете на улице статистика и спросите, что такое конфаундеры, вы, скорее всего, услышите самое запутанное и переусложненное объяснение, какое вам только доводилось слышать от ученого. Одна недавно вышедшая книга, написанная сразу двумя светилами статистики, объясняет, что это такое, на протяжении целых двух страниц, и мне еще, надеюсь, предстоит встретить ее читателя, который понял это объяснение.

Причина этих трудностей в том, что конфаундеры — понятие за рамками статистики. Это несоответствие того, что мы хотели бы получить (причинно-следственная связь), и того, что мы реально получаем статистическими методами. Если мы

не в состоянии математически выразить то, что собираемся найти, то как мы определим несоответствие ему? Исторически концепция конфаундеров возникла вокруг двух связанных между собой концепций: несопоставимости и скрытой (вмешивающейся) третьей переменной. Обе эти концепции упорно не поддавались формализации. Когда мы говорили о сопоставимости в контексте эксперимента Даниила, мы утверждали, что подопытная и контрольная группы должны быть идентичны по всем важным параметрам. Но из этого неизбежно следует, что нам придется отличать важное от неважного. Откуда мы знаем, что в исследовании про ходьбу пожилых мужчин в Гонолулу возраст — это важный параметр? Почему мы знаем, что расположение фамилий участников этого исследования по алфавиту — параметр неважный? Можно сказать, что это очевидно или что это следует из здравого смысла, однако бесчисленные поколения ученых бьются над тем, чтобы как-то формализовать этот здравый смысл, поскольку робота поступать согласно человеческому здравому смыслу научить нельзя.

От такой же двусмысленности страдает и определение третьей переменной. Считать ли таковой только общую причину X и Y, или достаточно, чтобы эта переменная была скоррелирована с ними обеими? Сегодня мы отвечаем на такие вопросы, обращаясь к каузальной диаграмме и выясняя, какие переменные отвечают за несоответствие между $P(X \mid Y)$ и $P(X \mid do(Y))$. Без диаграмм и оператора do пять поколений статистиков и медиков мучились с их суррогатами, ни один из которых не был полностью удовлетворяющим. То, что лекарства в вашей аптечке разработаны и испытаны на основе сомнительного определения конфаундеров, должно вызывать беспокойство.

Давайте взглянем на некоторые суррогатные дефиниции конфаундеров. Большинство их подпадает под одну из двух категорий — декларативную или процедурную. Типичное (и неверное) декларативное определение звучит так: «Конфаундер — это любая переменная, коррелирующая сразу и с X, и с Y». Процедурное определение, в свою очередь, будет пытаться определить конфаундер в терминах статистического

анализа. Это нравится статистикам, которые обожают методы, применимые на имеющихся данных напрямую, без обращения к модели.

Вот процедурное определение, известное под пугающим названием «несхлопываемость». Оно появилось в статье 1996 года норвежского эпидемиолога Свена Хернберга: «Формально можно сравнить грубый относительный риск и относительный риск после поправок на потенциальные конфаундеры. Наличие разницы означает, что конфаундеры реально присутствуют, и в этом случае следует использовать скорректированную оценку риска. Если разницы нет или она пренебрежимо мала, конфаундеров нет и предпочтительнее использовать грубую оценку». Другими словами, чтобы узнать, есть ли влияние конфаундеров, попробуйте вводить по ним поправки или не вводить; если есть разница, есть и конфаундер. Конечно, Хернберг был далеко не первым, кто предложил такой подход; почти столетие он путал эпидемиологов, экономистов, социологов и до сих царит в некоторых областях практической статистики. Я выбрал определение Хернберга только потому, что он написал об этом неожиданно подробно и в 1996 году, когда Революция Причинности уже шла полным ходом.

Самое популярное из декларативных определений образовалось за некоторый промежуток времени. Альфредо Морабиа, автор книги «История методов и концепций в эпидемиологии», называет его «классическим эпидемиологическим определением конфаундеров» и оно состоит из трех частей. Конфаундером X (экспериментального воздействия) и Y (результата) называется переменная Z, которая: 1) ассоциирована с X в популяции в целом и 2) ассоциирована с Y среди тех, кто не получал экспериментального воздействия X. В последние годы к этому добавилось третье условие: Z не должно находиться на каузальном пути от X к Y.

Обратите внимание, что вся терминология в классической версии (1 и 2) чисто статистическая. В частности, допускается только, что Z ассоциировано с X и Y, а не является причиной их обеих. Эдвард Симпсон в 1951 году предложил довольно невразумительное условие: «Y ассоциируется с Z среди неэкс-

понированного». С каузальной точки зрения похоже, что идеей Симпсона было исключить ту часть корреляции Z с X, которая возникает благодаря каузальному воздействию X на Y; другими словами, он хотел сказать, что Z воздействует на Y независимо от его воздействия на X. Единственное, что ему удалось придумать для выражения этого исключения, сосредоточив внимание на контрольной группе ($X = 0$), было введение поправок по X. Статистический словарь, лишенный слова «воздействие», не оставлял ему возможности сказать это иначе.

Вам кажется, что это все сбивает с толку? Так оно и есть. Насколько проще было бы, если бы он мог просто нарисовать каузальную диаграмму, вроде той, что на рис. 26, и сказать «Y ассоциирована с Z через пути, не проходящие через X». Но у него не было этого инструмента, и он не мог говорить о путях, концепция которых была тогда под запретом.

У «классического эпидемиологического определения» конфаундеров есть и другие недостатки, как показывают следующие два примера:

$$1)\ X \rightarrow Z \rightarrow Y$$

и

$$2)\ X \rightarrow \mathrm{M} \rightarrow Y$$
$$\downarrow$$
$$Z$$

В первом примере Z удовлетворяет условиям (1) и (2), но это не конфаундер. Такие переменные называют медиаторами или опосредующими переменными: они объясняют каузальное воздействие X на Y. Если вы пытаетесь определить каузальное воздействие X на Y, попытка вводить поправки по фактору Z приведет к неудаче. Если брать только тех индивидов как в контрольной, так и в опытной группе, для которых $Z = 0$, вы полностью блокируете воздействие X, потому что оно работает посредством изменения Z. Из этого вы делаете неверный вывод, что X не влияет на Y. Именно это имел в виду Эзра Кляйн, когда говорил: «Иногда в итоге вы выравниваете выборку как раз по тому фактору, который хотите измерить».

Во втором примере Z — это опосредованная переменная для медиатора M. Статистики очень часто используют опосредо-

ванные переменные, когда истинная каузальная переменная не поддается измерению: так, принадлежность к политической партии может быть использована как опосредованная переменная для политических взглядов. Поскольку Z не является точной мерой M, некоторая часть влияния X на Y способна просочиться, если вы вводите поправки по Z. Тем не менее это все еще ошибочно; хотя смещение будет меньшим, чем если вы вводите поправки по M, оно никуда не денется.

По этой причине позднее статистики, среди которых стоит отметить Дэвида Кокса с его учебником «Планирование исследований» (1958), предупреждали, что вводить поправки по Z стоит только в том случае, если вы «заранее имеете серьезные причины предполагать», что на Z не влияет X. Эти «заранее известные серьезные причины» — не что иное, как каузальное допущение. Он добавляет: «Выдвигать такие гипотезы совершенно нормально, однако ученый должен четко осознавать, когда именно к ним апеллировать». Напомню, что это 1958 год, разгар запрета на обсуждение причинности. Кокс открыто говорит, что при введении поправок по конфаундерам вполне допустимо украдкой глотнуть запретного — главное, не говорить об этом святошам. Дерзкое предложение! Я никогда не упускаю случая отдать должное его храбрости.

К 1980 году условия Симпсона и Кокса были объединены в трехчастную проверку на конфаундеры, упомянутую выше. Она примерно настолько же надежна, как лодка, которая течет всего в трех местах. Хотя она и обращается нерешительно к причинности в третьей части, несложно показать, что каждая из первых двух и не нужна, и недостаточна. Гренланд и Робинс вынесли это вердикт в своей эпохальной статье 1986 года. Они сформировали совершенно новый подход к проблеме конфаундеров, который назвали взаимозаменяемостью. Они вернулись к исходной идее о том, что контрольная группа ($X = 0$) должна быть сравнима с опытной группой ($X = 1$). Однако они добавили к ней контрфактивный выверт (вспомним из главы 1, что контрфактивные высказывания находятся на третьей ступени Лестницы Причинности и поэтому обладают достаточной мощностью, для того чтобы распознавать конфаундеры).

Взаимозаменяемость требует от исследователя рассмотреть опытную группу, вообразить, что стало бы с составляющими ее объектами, если бы изучаемое воздействие не применялось, и затем решить, будет ли результат таким же, как и для тех, кто не подвергался (в реальности) этому воздействию. Только в случае положительного ответа мы можем сказать, что в исследовании нет конфаундеров.

В 1986 году говорить с эпидемиологической аудиторией о контрфактивных высказываниях было достаточно смело, потому что они в значительной степени оставались под влиянием классической статистики, полагающей, что все ответы уже находятся в данных, а не в том, что могло произойти и навеки останется ненаблюдаемым.

Однако статистическое сообщество было частично подготовлено к подобной ереси, за что стоит благодарить пионерскую работу другого статистика из Гарварда, Дональда Рубина. В рубинской схеме потенциальных исходов, предложенной в 1974 году, контрфактивные переменные вроде «артериальное давление испытуемого *X*, если бы он получал препарат *П*» и «артериальное давление испытуемого *X*, если бы он не получал препарата *П*» столь же легитимны, как традиционные переменные вроде артериального давления — несмотря на тот факт, что наблюдения за одной из этих переменных не состоятся никогда.

Робинс и Гренланд решили выразить свою концепцию конфаундеров в терминах потенциальных исходов. Они разделили выборку на четыре типа испытуемых: обреченных, каузативных, превентивных и иммунных. Давайте представим, что экспериментальное воздействие *X* — это вакцина от гриппа, а исход *Y* — заболевание гриппом. Обреченные — это те, кому вакцина не помогает, они заболеют гриппом вне зависимости от того, получат вакцину или нет. Каузативная группа (которой в реальности может не быть вовсе) включает тех, у кого вакцина вызывает настоящий грипп. Превентивная группа состоит из тех, для кого вакцина предотвращает заболевание: они заболеют гриппом, если не привьются, и не заболеют, если сделают прививку. Наконец, иммунная группа — это те, кто

не заболеет гриппом ни в каком случае. Табл. 4 суммирует эти соображения.

Таблица 4

Группа	Процент	Исход, если вакцинированы	Исход, если не вакцинированы
Обреченные	О	Грипп	Грипп
Каузативные	К	Грипп	Нет гриппа
Превентивная	П	Нет гриппа	Грипп
Имунная	И	Нет гриппа	Нет гриппа

В идеале у каждого человека на лбу должна быть этикета, сообщающая, к какой группе он принадлежит. Взаимозаменяемость предполагает, что процент людей с каждым типом этикетки (процент *О*, процент *К*, процент *П* и процент *И* соответственно) должен быть одинаков и в контрольной, и в опытной группе. Равенство этих пропорций гарантирует, что исход будет тем же самым, если мы поменяем местами опыт и контроль. В противном случае опытная и контрольная группа неодинаковы и наши оценки эффективности вакцины окажутся смещенными. Обратите внимание, что две группы могут различаться по самым разным параметрам: по возрасту, полу, состоянию здоровья и ряду других характеристик. Только наличие равенства по процентному соотношению *О*, *К*, *П* и *И* определяет, взаимозаменяемы они или нет. Таким образом, взаимозаменяемость сводится к равенству между двумя наборами из четырех пропорций, что намного проще альтернативы — учета бесчисленных факторов, по которым популяции могут различаться.

Используя это определение конфаундеров, опирающееся на здравый смысл, Гренланд и Робинс показали, что статистические определения, как декларативные, так и процедурные,

дают неверные ответы. Переменная может удовлетворять трехчастному тесту эпидемиологов и все-таки усиливать смещение оценки, если вносить в нее поправку.

Определение Гренланда и Робинса было огромным достижением, потому что оно позволило им привести ясные примеры, наглядно демонстрирующие, что предыдущие определения были неадекватны. Тем не менее эту дефиницию нельзя перевести в практическую плоскость. Проще говоря, таких удобных этикеток на лбу не бывает. Мы даже не можем подсчитать процент *О*, *К*, *П* и *И*. Это как раз та информация, которую хитрый джинн природы прячет от всех внутри своей волшебной лампы. Без этой информации исследователю остается только полагаться на интуицию, решая, взаимозаменяемы опытная и контрольная группа или нет.

Надеюсь, к этому моменту мне удалось разжечь ваше любопытство. Каким образом каузальные диаграммы превращают головную боль конфаундеров в веселую игру? Секрет лежит в операционном тесте на конфаундеры, называемом критерием черного хода. Этот критерий превращает проблему определения конфаундеров, их поиска и ввода поправок по ним в рутинную задачу, ничуть не более сложную, чем решение журнальной головоломки. Он привел столетнюю, упорную проблему к благополучному разрешению.

Оператор *Do* и критерий черного хода

Чтобы понять, как работает критерий черного хода, лучше сначала интуитивно представить себе, как двигается информация в каузальной диаграмме. Мне нравится представлять связи как трубы, по которым информация распространяется от стартовой точки *X* до финиша *Y*. Не забывайте, что распространение информации идет одновременно по двум направлениям — по каузальному и некаузальному, как мы видели в главе 3.

На самом деле некаузальные пути как раз и являются источником конфаундеров. Вспомним, что я определяю их как все,

что вынуждает $P(Y \mid do(X))$ отличаться от $P(Y \mid X)$. Оператор *do* стирает все стрелки, которые входят в X и предотвращает движение информации от X в некаузальном направлении. Таким же эффектом обладает рандомизация. Наконец, к тому же самому приводит введение статистических поправок, если правильно выбрать переменные, по которым эти поправки следует вводить.

В предыдущей главе мы рассмотрели три правила, которые рассказывают нам, как остановить поток информации по любому отдельно взятому соединению. Я повторю их, чтобы подчеркнуть:

а) в соединении типа «цепочка» $A \rightarrow B \rightarrow C$ введение поправок по B предотвращает движение информации об A к C и наоборот;

б) в вилке, или вмешивающемся соединении $A \leftarrow B \rightarrow C$ поправки по B также предотвращают движение информации об A к C и наоборот;

в) в коллайдере $A \rightarrow B \leftarrow C$ действуют прямо противоположные правила. Переменные A и C изначально независимы, поэтому информация об A ничего не говорит о C. Но если вы вводите поправки по B, информация начинает распространяться по «трубе», благодаря эффекту объяснения. Мы должны также держать в уме еще одно фундаментальное правило:

г) выравнивание по нисходящей или опосредованной переменной подобно частичному выравниванию по исследуемой переменной. Выравнивание по переменной, нисходящей по отношению к медиатору, частично закрывает трубу; выравнивание по переменной, нисходящей по отношению к точке схождения, частично открывает трубу.

А что же будет в случае более длинных труб с большим числом соединений, вроде такой: $A \leftarrow B \leftarrow C \rightarrow D \leftarrow E \rightarrow F \rightarrow \rightarrow G \rightarrow H \leftarrow I \leftarrow J$?

Ответ очень прост: если хоть одна связь окажется заблокирована, то J ничего не сможет «узнать» про A по этому пути. Таким образом, у нас множество вариантов прервать сообщение между A и J: вводить поправки по B, по C, не вводить

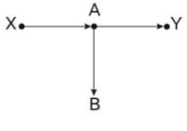
поправки по *D* (потому что это коллайдер), вводить по *E* и т.д. Достаточно любого из этих вариантов.

Вот почему обычная статистическая процедура выравнивания по всем параметрам, которые только можно измерить, так ошибочна. На самом деле приведенный выше путь заблокирован даже в том случае, если мы не вводим никаких поправок! Коллайдеры к *D* и *G* закрывают путь без посторонней помощи. Введение поправок по *D* и *G* откроет этот путь и позволит *J* «услышать» *A*.

Итак, чтобы устранить конфаундеры между *X* и *Y*, нам необходимо только заблокировать все некаузальные пути между ними, не блокируя и не нарушая каузальные пути. Выражаясь точнее, путь черного хода — это любой путь от *X* до *Y*, который начинается со стрелки, входящей в X. Конфаундеры между *X* и *Y* будут устранены, если мы закроем все черные ходы (потому что такие пути допускают ложную корреляцию между *X* и *Y*). Если мы делаем это, выравнивая выборку по некоторому набору переменных *Z*, следует также убедиться, что ни один фактор из *Z* не является нисходящей переменной по отношению к *X* на каузальном пути, иначе этот путь полностью или частично закроется.

Вот и все! С этими правилами устранение конфаундеров становится настолько элементарным делом, что можно воспринимать его как игру. Я предлагаю вам несколько примеров, чтобы войти во вкус и увидеть, как это просто. Если вам все еще кажется, что это сложно, будьте уверены, что существуют алгоритмы, решающие все эти задачи в течение наносекунд. В каждом случае цель игры — определить набор переменных, которые устранят конфаундеры между *X* и *Y*. Другими словами, они не должны исходить от *X* и они должны блокировать все черные ходы.

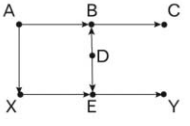
Игра 1



Эта — совсем простая! К X не идет ни одной стрелки, следовательно, черных ходов нет. Нам не нужно вводить никаких поправок.

Тем не менее некоторые исследователи сочтут B конфаундером. Оно связано с X по цепочке $X \rightarrow A \rightarrow B$. Оно связано с Y у особей, у которых $X = 0$, потому что имеется открытый путь $B \leftarrow A \rightarrow Y$, не проходящий через X. И при этом B не находится на каузальном пути $X \rightarrow A \rightarrow Y$. Таким образом, оно проходит трехступенчатое «классическое эпидемиологическое» определение конфаундера, но не соответствует критерию черного хода и поправки, введенные по нему, чреваты неприятностями.

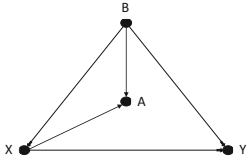
Игра 2



В этом примере следует рассматривать A, B, C и D как «доэкспериментальные» переменные (экспериментальное воздействие, как всегда, обозначено X). Теперь имеется один черный ход $X \leftarrow A \rightarrow B \leftarrow D \rightarrow E \rightarrow Y$. Этот путь уже блокирован коллайдером в B, поэтому нам опять не нужно вводить никаких поправок. Многие статистики стали бы выравнивать выборки по B или C, думая, что в этом нет вреда, поскольку они случаются до опыта. Один известный статистик еще совсем недавно писал: «Избегание введения поправок по некоторым наблюдаемым ковариантам… это ненаучная кустарщина». Он неправ: поправки по B или C — плохая идея, потому что они откроют некаузальный путь и создадут конфаундеры между X и Y. Обратите внимание, что в этом случае мы можем снова закрыть этот путь, корректируя по A или D. Этот образец показывает, что доступны различные стратегии устранения конфаундеров. Одни исследователи пойдут легким путем и не будут вводить никаких поправок; более традиционный подход предполагает корректировку по C и D. Оба варианта верны и приведут

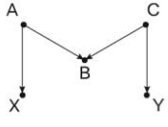
к одному и тому же результату (если модель верна, а выборка достаточно велика).

Игра 3



В играх 1 и 2 вам не нужно было ничего делать, но теперь придется. Имеется один черный ход от X к Y, $X \rightarrow B \leftarrow Y$, который можно заблокировать, только вводя поправки по B. Если B невозможно наблюдать непосредственно, тогда оценить влияние X на Y невозможно без проведения рандомизированного контролируемого исследования. Некоторые (на самом деле почти все) статистики в этой ситуации будут выравнивать по A, как по опосредованной переменной для не поддающейся наблюдению переменной B, но это только частично устраняет смещение от конфаундера и вносит новое смещение от схождения.

Игра 4



Эта игра представляет новый для нас тип смещения оценки — M-тип (названный так по форме данного графа). Снова у нас только один черный ход, уже заблокированный коллайдером в B. Таким образом, нам снова не нужно вводить поправки. Тем не менее все статистики до 1986 года и многие даже сегодня посчитали бы B конфаундером. B ассоциировано с X (посредством $X \leftarrow A \rightarrow B$) и с Y через путь, который не проходит через X ($B \rightarrow C \leftarrow Y$). Оно не лежит на каузальном пути и не является

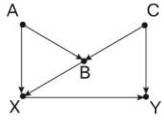
нисходящим по отношению к чему-либо на каузальном пути, потому что от *X* к *Y* каузального пути нет. Таким образом, *B* проходит традиционный трехступенчатый тест на конфаундеры.

М-тип смещения оценки показывает пальцем, что неверно в традиционном подходе. Неверно называть такую переменную, как *B*, конфаундером только потому, что она ассоциирована сразу и с *X*, и с *Y*. Повторяем, между *X* и *Y* нет вмешивающихся переменных, если мы не вводим поправки по *B*. *B* становится осложнителем только тогда, когда мы корректируем данные по нему!

Когда в 90-х годах XX века я начал показывать эту диаграмму статистикам, некоторые из них смеялись и говорили, что на практике вероятность столкнуться с такой схемой ничтожно мала. Я не согласен! Так, использование ремней безопасности в автомобиле (*B*) не влияет каузально ни на курение (*X*), ни на рак легких (*Y*), это просто показатель отношения индивида к соблюдению общественных норм (*A*) и мер безопасности и охраны здоровья (*C*). Образ жизни, вытекающий из этого отношения, может влиять на подверженность заболеваниям легких (*Y*). На практике соблюдение использования ремней безопасности оказалось скоррелировано и с *X*, и с *Y*. В исследовании 2006 года этот фактор значился одним из первых в списке переменных, по которым предполагалось вводить поправки. Если принять приведенную выше модель, то введение поправок только по *B* будет ошибочно.

Обратите внимание, что вводить поправки по *B* совершенно нормально, если при этом корректировать результаты также по *A* или *C*. Корректировка открывает трубу коллайдеру *B*, точке схождения, но дополнительная корректировка по *A* или *C* закрывает ее снова. К сожалению, в примере с ремнями безопасности и курением *A* и C — переменные, связанные с отношением людей к определенным вопросам, и получить данные по ним крайне сложно. А если переменная ненаблюдаема, по ней невозможно внести поправку.

Игра 5



Игра 5 — почти то же самое, что и игра 4, но с небольшим дополнительным вывертом. Теперь требуется закрыть второй черный ход $X \to B \to C \leftarrow Y$. Если мы блокируем этот путь, вводя поправки по B, у нас открывается M-образный путь $X \to A \leftarrow B \to C \leftarrow Y$. Чтобы закрыть этот путь, нам придется вводить поправки также по A или C. Однако обратите внимание, что мы не сможем обойтись поправками только по C, это закроет путь $X \to B \to C \leftarrow Y$, но не затронет второй.

Игры 1—3 взяты из статьи 1993 года под названием «В поисках более ясного определения конфаундеров», написанной Кларис Вайнберг, заместителем начальника Национальных институтов здравоохранения. Она вышла в переходный период между 1986 и 1995 годами, когда статья Гренланда и Робинса уже была доступна, но о каузальных диаграммах еще не было широко известно. Поэтому Вайнберг была вынуждена немало потрудиться, арифметически доказывая взаимозаменяемость в каждом из приведенных примеров. Хотя для передачи обсуждаемых сценариев она и использовала графику, логика диаграмм для различения конфаундеров и переменных, устраняющих осложнения, не применялась. Кроме нее я не знаю никого, кому бы это удалось. Позже, в 2012 году, она стала соавтором дополненной версии статьи, где те же примеры проанализированы с помощью каузальных диаграмм и подтверждено, что все ее выводы 1993 года верны.

В обеих статьях Вайнберг медицинское применение приведенных схем было в выяснении влияния курения (X) на выкидыши, или «спонтанное прерывание беременности» (Y). В игре 1 фактор A — это нарушения, вызываемые курением; это ненаблюдаемая переменная, потому что мы не знаем, в чем эти нарушения состоят. Фактор B представляет собой историю предыдущих выкидышей. Для эпидемиолога будет большим искушением обратить внимание на число предшествовавших

выкидышей и ввести поправку по этой переменной, оценивая вероятность будущих выкидышей. Но в данном случае это как раз неправильно! Поступив так, мы частично деактивируем механизм, по которому действует курение, и, таким образом, истинное влияние курения окажется недооцененным.

Игра 2 — более сложная версия, в которой курение разделено между двумя разными переменными: *X* показывает, курит ли беременная сейчас (в начале второй беременности), а *A* — курила ли она во время первой беременности. *B* и *E* — скрытые (ненаблюдаемые) нарушения развития, вызываемые курением, *D* — физиологические причины этих нарушений. Обратите внимание, что эта диаграмма допускает вариант, при котором женщина изменила свое поведение между беременностями, начав или бросив курить, но другие физиологические причины нарушений остаются прежними. Многие эпидемиологи захотят ввести поправки по предшествовавшим случаям выкидышей, но это плохая идея, если одновременно не ввести поправки по курению в предыдущей беременности (*A*).

Игры 4 и 5 взяты из статьи, опубликованной в 2014 году Эндрю Форбсом, биостатистиком из Университета Монаша в Австралии, вместе с несколькими соавторами. Он интересовался влиянием курения на развитие астмы у взрослых. В игре 4 переменная *X* — это отношение индивида к курению, *Y* — болел ли он астмой во взрослом возрасте. Фактор *B* обозначает заболевание астмой в детском возрасте, и это фактор-коллайдер, потому что на него влияют одновременно *A*, курение родителей и *C* — скрытая (и ненаблюдаемая) предрасположенность к астме. В игре 5 у переменных те же значения, но Форбс добавил две стрелки для пущего реализма (смысл игры 4 был только в том, чтобы представить читателям *M*-образный граф).

На самом деле в полной модели Форбса было еще несколько переменных и она выглядела так, как на рис. 27. Обратите внимание, что игра 5 «погружена» в эту модель, в том смысле, что у переменных *A*, *B*, C, *X* и *Y* ровно те же взаимоотношения. Следовательно, наши рассуждения остаются верными и в этом случае, и нам следует ввести поправки по *A* и *B* или по *C*, но *C* — ненаблюдаемая и, следовательно, невыравниваемая

переменная. Вдобавок к этому у нас появляются еще четыре вмешивающиеся переменные: *D* — астма у родителей, *E* — хронический бронхит, *F* — пол и *G* — социально-экономический статус. Читатель для своего удовольствия может самостоятельно вывести, что нам нужно вводить поправки по *E*, *F* и *G*, но не нужно по *D*. Таким образом, достаточный набор переменных для устранения смещения оценки — это *A*, *B*, *E*, F и *G*.

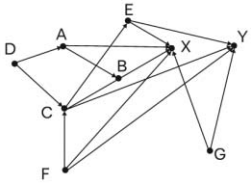


Рис. 27. Модель связи курения (*X*) и астмы (*Y*), созданная Эндрю Форбсом

В конце концов Форбс обнаружил, что в необработанных данных у курения есть небольшая и статистически незначимая ассоциация с астмой среди взрослых, а после введения поправок по конфаундерам этот эффект становился еще незаметнее и менее достоверным. Нулевой результат не должен, однако, уводить внимание от того факта, что его статья — модельная работа по «искусному допросу Природы».

И последний комментарий по поводу этих игр: когда переменными становятся такие явления, как курение, выкидыши и т.д., это уже совершенно определенно не игры, а серьезное дело. Я назвал их играми, потому что радость от возможности получить их решение быстро и осмысленно очень похожа на удовольствие, которое ребенок получает, обнаружив, что щелкает как орехи те задачи, что раньше вводили его в ступор.

Мало какие моменты в научной карьере приносят столько удовлетворения, как удачное сведение проблемы, озадачивавшей и путавшей целые поколения предшественников, к прямолинейной игре или алгоритму. Я полагаю окончательное решение вопроса конфаундеров одним из важнейших событий Революции Причинности, потому что оно положило конец эпохе

заблуждений, приводивших, вероятно, к множеству неверных решений в прошлом. Это была тихая революция, баррикады которой располагались в основном в исследовательских лабораториях и конференц-залах. Однако, вооружившись новыми инструментами и идеями, научное сообщество сегодня берется за более сложные задачи, как теоретические, так и практические, о чем расскажут последующие главы.

Глава 5
Дымные дебаты: на свежий воздух

И сказали друг другу:
«пойдем бросим жребии, чтобы узнать,
за кого постигает нас эта беда».
Книга пророка Ионы. 1:7

В конце 50-х — начале 60-х годов XX века статистика и медицина бились над одним из самых важных вопросов столетия в области здравоохранения: вызывает ли курение рак легких? Спустя полстолетия мы воспринимаем ответ на этот вопрос как само собой разумеющееся, но в те времена он был совершенно неочевиден. Ученые и даже члены одной семьи часто имели по этому вопросу противоположные мнения.

Семья Якоба Ерушалми (1904—1973) была одной из таких. Биостатистик Калифорнийского университета в Беркли Ерушалми был одной из последних цитаделей одобрения курения в академическом мире. «Ерушалми возражал против мнения, что сигареты вызывают рак, вплоть до своего последнего дня», — писал его племянник Дэвид Лилиенфельд спустя много лет. Отец Дэвида Эйб Лилиенфельд был эпидемиологом в Университете Джонса Хопкинса и одним из самых ярых сторонников теории, что курение рак все-таки вызывает. Лилиенфельд вспоминает, как дядя Як (сокращенно от Якоб) с его отцом сидели и спорили о последствиях курения, окутанные «облаками дыма от сигарет

Яка и трубки Эйба» (см. иллюстрацию в начале главы). Ах, если бы Революция Причинности могла привнести в их спор глоток свежего воздуха! Как будет видно из этой главы, одним из самых серьезных аргументов против гипотезы, что курение вызывает рак, было вероятное наличие неизвестных факторов, которые обусловливают одновременно тягу к никотину и рак легких. Мы только что обсудили подобные осложняющие схемы и отметили, что современные каузальные диаграммы извели проклятье конфаундеров под корень. Но сейчас мы с вами в 1950-х и 1960-х, за два десятилетия до Сандера Гренланда и Джейми Робинса и за 30 лет до того, как хоть что-нибудь будет известно об операторе *do*. Поэтому интересно посмотреть, как ученые той эпохи справились с этой проблемой и доказали, что аргумент осложнения непрочен как дым.

Нет сомнений, что темой многих споров Эйба и Яка в курилке был не табак и не рак, а это несносное слово «причина». Не в первый раз в истории врачи сталкивались со сложными каузальными вопросами: многие великие вехи в истории медицины — это обнаружение истинных каузативных агентов. В середине *X*VIII века Джеймс Линд открыл, что плоды цитрусовых способны предотвращать цингу, а в середине *XIX*-го Джон Сноу выяснил, что вода, загрязненная испражнениями, вызывает холеру (позже исследования более точно идентифицировали в каждом случае каузативные агенты: дефицит витамина С для цинги, холерный вибрион для холеры). Эти два примера блестящего расследования объединяет то, что соотношение между причиной и следствием в обоих случаях однозначно. Холерный вибрион — единственная причина холеры или, как мы сказали бы сегодня, ее необходимое и достаточное условие. Если вы не подверглись его воздействию, вы не заболеете холерой. Точно так же недостаток витамина *C* — необходимое условие для развития цинги, а по прошествии некоторого времени — и достаточное.

Спор о курении и раке поставил под сомнение эту монолитную концепцию причинности. Многие люди курят всю свою жизнь и не заболевают раком легких. И наоборот, этим недугом страдают те, кто за всю жизнь не сделал ни одной

затяжки. У одних он развивается из-за наследственной предрасположенности, у других — под воздействием вызывающих рак веществ, у третьих — по обеим причинам.

Конечно, к этому времени статистике уже был известен превосходный способ установления причинности в более широком смысле: рандомизированное контролируемое исследование. Но такое исследование в случае курения и невозможно, и неэтично. Как прикажете предписать случайно выбранным людям курить в течение десятилетий, вероятно нанеся ущерб их здоровью, только для того, чтобы посмотреть, будет ли у них в итоге рак легких? Нельзя себе представить, чтобы хоть где-нибудь, кроме Северной Кореи, кто-то «добровольно вызвался» быть испытуемым в подобном исследовании.

Без РКИ не было шансов убедить скептиков, таких как Ерушалми или Р. Э. Фишер, которые были приверженцами идеи, что наблюдаемая ассоциация между курением и раком легких ложная. По их мнению, ассоциацию порождал некий скрытый третий фактор. Например, мог существовать «ген курильщика», благодаря которому у людей была бы тяга к табаку и одновременно бóльшая вероятность заболеть раком легких (скорее всего, в результате других особенностей в выборе образа жизни). Предлагаемые ими конфаундеры были в лучшем случае неправдоподобны. Однако именно на противников курения возлагалось бремя доказательств, что конфаундеров в этом случае нет, т.е. доказать отсутствие чего-либо, что практически невозможно, Фишер и Ерушалми хорошо знали об этом.

Финальный прорыв из сложившейся патовой ситуации — история одновременно и о великом триумфе, и об упущенной возможности. Это был триумф для здравоохранения, потому что в конце концов эпидемиологи докопались до истины. Отчет начальника медицинской службы Соединенных Штатов Америки в 1964 году недвусмысленно гласил: «Курение является причиной рака легких у взрослых мужчин». Это утверждение навсегда уничтожило аргумент, что влияние курения на заболеваемость раком легких не доказано. Доля курящих мужчин в США уже на следующий год начала снижаться, и к настоящему времени составляет меньше половины той, которая была

в 1964 году. Нет сомнений в том, что в результате миллионы жизней удалось спасти, а продолжительность жизни выросла.

Однако триумф был неполным. Промежуток времени, потребовавшийся для того, чтобы прийти к вышеупомянутому заключению, примерно с 1950 по 1964 годы, мог бы быть значительно короче, если бы ученые воспользовались более строгой теорией причинности. И, что особенно важно с точки зрения этой книги, исследователи 60-х годов XX века так и не сумели собрать такую теорию воедино. Чтобы подтвердить мнение о том, что курение вызывает рак, комитет начальника медицинской службы воспользовался серией неформальных руководящих принципов, известных как критерии Хилла, названные по имени статистика из Лондонского университета Остина Бредфорда Хилла. У каждого из этих критериев есть исключения, хотя все вместе они весьма убедительны, апеллируют к здравому смыслу и даже мудрости. Из избыточно методологичного мира Фишера принципы Хилла переносят нас в противоположную реальность — мир без методологии, где причинность выявляется на основе количественных паттернов статистических тенденций. Революция Причинности выстраивает мост между двумя этими крайностями, вооружая наше интуитивное чувство причинно-следственных связей мощью математического аппарата. Но проделать эту работу предстояло уже следующему поколению.

Табак: рукотворная эпидемия

В 1902 году сигареты занимали только 2% на рынке табака в Соединенных Штатах Америки — вездесущим символом потребления этого продукта были плевательницы, а не пепельницы. Но американские привычки изменились под воздействием двух влиятельных сил, работавших рука об руку: рекламы и автоматизации. Производимые машинами сигареты благодаря цене и доступности легко вытеснили как вручную изготовляемые сигары, так и курительные трубки. Одновременно с этим табачная промышленность изобрела и довела

до совершенства многие рекламные приемы. Люди, смотревшие телевизор в 60-е годы XX века, легко вспомнят любое количество цепких рекламных роликов — от «Вам многое понравится в «Мальборо»» до «Это был долгий путь, беби».

К 1952 году доля сигарет на табачном рынке взлетела вверх с 2 до 81%, да и сам рынок очень вырос. Эта кардинальная смена привычек целой страны имела неожиданные последствия для здоровья нации. Даже в первые годы XX столетия имелись подозрения, что курение нездорово, что оно «раздражает» горло и вызывает кашель. К середине века накапливающаяся информация становится все более зловещей. До распространения сигарет рак легких встречался так редко, что врачу обычно случалось сталкиваться с ним лишь однократно на протяжении всей своей практики. Однако за период с 1900 до 1950 года когда-то редкое заболевание стало встречаться вчетверо чаще, а к 1960 году оно стало самой распространенной формой рака среди мужчин. Столь значительное изменение в распространенности смертельного заболевания нуждалось в объяснении.

Задним умом легко обвинить в этом курение. Если мы наложим графики распространения рака легких и потребления табака, не заметить связи невозможно. Однако графики изменений по времени очень плохо доказывают причинно-следственные связи. Между 1900 и 1950 годами стало иным очень многое, что также могло быть сочтено причиной перемен: асфальтирование дорог, автомобильный выхлоп, содержащий свинец, общее загрязнение воздуха. Британский эпидемиолог Ричард Долл писал в 1991 году: «Автомобилизация была новым фактором, и, если бы я жил в те времена, я бы, скорее всего, поставил все на выхлопные газы или асфальтирование дорог» (рис. 28).

Задача науки — отставить догадки в сторону и приглядеться к фактам. В 1948 году Долл и Хилл объединили усилия в поисках причин эпидемии рака. Хилл был главным по статистике в очень успешном рандомизированном контролируемом эксперименте, результаты которого были опубликованы ранее в том же году; было доказано, что стрептомицин — один из первых антибиотиков — эффективен против возбудителя туберкулеза. Эта

работа, одна из самых значительных вех в истории медицины, не только познакомила врачей с «чудо-лекарствами», но и подкрепила репутацию РКИ, которые вскоре стали стандартом клинических исследований в эпидемиологии.

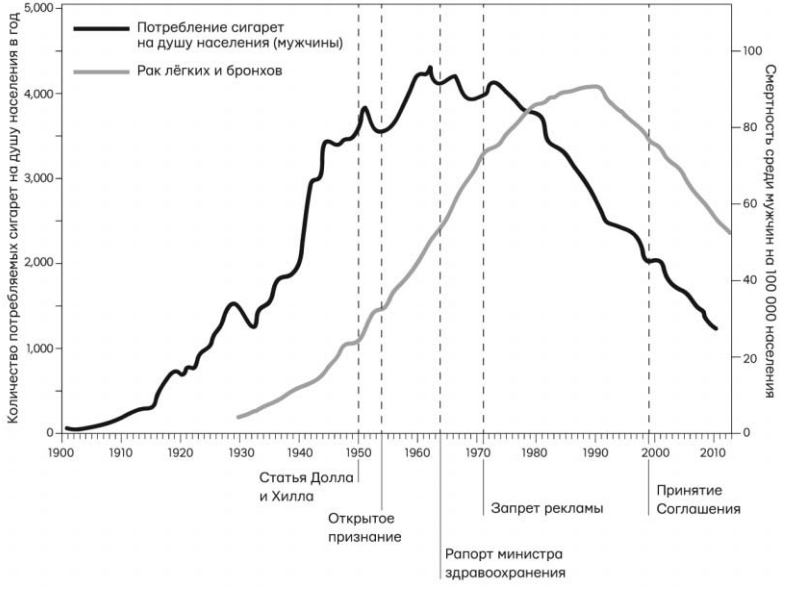


Рис. 28. Графики потребления сигарет на душу населения в США (черным) и смертности от рака легких среди мужчин (серым) удивительным образом совпадают: кривая раковых заболеваний практически повторяет кривую курения с запаздыванием примерно в 30 лет. Тем не менее это свидетельство не доказывает причинно-следственную связь. На графике отмечены некоторые ключевые даты, в том числе дата публикации статьи Ричарда Долла и Остина Бредфорда Хилла в 1950 году, которая первой привлекла внимание многих профессионалов мира медицины к ассоциации между курением и раком легких (источник: график Маян Харел с использованием данных Американского общества по борьбе с раком, Центра борьбы с заболеваниями и архива Министерства здравоохранения США)

Конечно, Хилл понимал, что РКИ в этом случае провести невозможно, но он знал о преимуществах сравнения опытной группы с контрольной группой. Поэтому он предложил сравнить

пациентов, у которых уже диагностировали рак, с контрольной группой здоровых волонтеров. В каждой группе участников опрашивали, выясняя их привычки в прошлом и истории болезней. Чтобы не создавать системной ошибки, проводящим опрос не говорили, у кого из испытуемых есть рак, а у кого нет.

Результаты исследования были шокирующими: из 649 опрошенных пациентов с раком легких все, кроме двоих, курили. Случайность этого была настолько статистически невероятна, что Долл и Хилл не смогли удержаться от того, чтобы ее не подсчитать; 1,5 миллиона к одному. Кроме этого, пациенты с раком легких в среднем были более заядлыми курильщиками, чем входящие в контрольную группу, но (противоречие, которое впоследствии очень педалировал Фишер) меньший процент из них при этом вдыхал дым.

Этот тип исследования сегодня называется «случай — контроль», потому что в нем сравниваются «случай» (люди с заболеванием) и контрольная группа. Это явный прогресс по сравнению с серийными данными за определенное время, потому что специалисты могут вносить поправки по конфаундерам, таким как возраст, пол и испытанное воздействие загрязнений окружающей среды. Тем не менее в дизайне «случай — контроль» есть некоторые очевидные недостатки. Он ретроспективен, это значит, что мы изучаем людей, о которых известно, что у них рак, и обращаем взгляд в прошлое, пытаясь понять, почему он возник. Логика вероятностей также направлена в противоположную сторону. Данные сообщают нам вероятность того, что больные раком в прошлом курили, хотя нас интересует вероятность того, что курящий заболеет раком. Именно эта последняя вероятность имеет значение для человека, которому решать, курить ему или нет.

Кроме этого, в исследованиях типа «случай — контроль» возможно еще несколько источников искажений. Один из них — «ошибка памяти». Хотя Долл и Хилл позаботились о том, чтобы опрашивающие ничего не знали о диагнозах испытуемых, сами пациенты знали наверняка, есть у них рак или нет. Это так или иначе влияло на их воспоминания. Другая проблема — ошибка отбора. Госпитализированные пациенты, больные раком, нико-

им образом не должны считаться репрезентативной выборкой всей популяции или даже курящей ее части.

Короче говоря, результаты Долла и Хилла были очень наглядны, но все же не могли считаться строгим доказательством того, что курение вызывает рак. Двое исследователей вначале осторожно называли корреляцию связью. Устранив несколько конфаундеров, они осмелились на более уверенное утверждение: курение является фактором, и важным фактором в развитии карциномы легкого.

За несколько последующих лет 19 исследований «случай — контроль», проведенных в различных странах, пришли практически к тому же заключению. Однако, как злорадно указал Р. Э. Фишер, эксперименты с искаженной оценкой ничего не доказывают, даже если повторить их 19 раз, — оценка остается искаженной. В 1957 году Фишер писал, что эти исследования «лишь повторяли свидетельства одного и того же типа, и необходимо попытаться выяснить, достаточно ли свидетельств этого типа для научного заключения».

Долл и Хилл понимали, что, если в испытаниях «случай — контроль» есть скрытая системная ошибка, одно только повторение исследований ее не устранит. Поэтому в 1951 году они начали эксперимент на перспективу, для которого разослали опросники 60 тысячам британских врачей с просьбой рассказать об их отношении к курению и стали наблюдать за ними в дальнейшем (Американское общество по борьбе с раком начало похожее и более масштабное исследование примерно в то же время). Всего за пять лет обнаружились драматические различия. Смертность курильщиков от рака легких в 24 раза превышала таковую у некурящих. В эксперименте Американского общества по борьбе с раком результаты были еще страшнее: курильщики умирали от рака легких в 29 раз чаще, чем некурящие, а те, кто курил много, — даже в 90 раз. В свою очередь, люди, которые бросили курить, снизили для себя риск заболевания вдвое. Согласованность всех этих результатов (большее потребление табака ведет к большему риску — прекращение курения снижает риск) была еще одним сильным аргументом в пользу причинно-следственной связи.

Врачи называют это эффектом «доза — ответ»: если вещество *A* вызывает биологический эффект *B*, то бóльшая доза *A* (как правило, хоть и не всегда) вызывает более сильный ответ *B*.

Тем не менее скептиков, таких как Фишер и Ерушалми, это не убедило. Исследования все-таки не способны были сравнить две во всем остальном идентичные группы, различающиеся только привычками по отношению к курению. На самом деле неясно, возможно ли такое исследование вообще. Группа курильщиков назначает себя сама. Курильщики могут генетически или «конституционально» отличаться от некурящих по ряду параметров — быть более склонными к риску, к злоупотреблению алкоголем. Некоторые из этих особенностей поведения способны оказывать негативное воздействие на здоровье, которое ошибочно приписывается курению. Для скептиков это был особенно удобный аргумент, потому что конституциональная гипотеза была практически непроверяемой. Только после секвенирования полного человеческого генома в 2000 году стало возможно пытаться найти гены, связанные с раком легких (по иронии судьбы Фишер оказался прав, хотя и в очень ограниченном смысле: такие гены существуют). Несмотря на это, в 1959 году Джером Корнфилд в соавторстве с Эйбом Лилиенфельдом опубликовал работу, пункт за пунктом опровергающую аргументы Фишера, которая в глазах многих врачей окончательно решила вопрос. Корнфилд, который работал в Национальном институте здоровья, был необычным участником спора о курении и раке. Далекий и от статистики, и от биологии по первому образованию, он специализировался на истории и изучил статистику в Департаменте сельского хозяйства США. Когда-то он выкуривал две с половиной пачки в день, но бросил курить, когда увидел данные по раку легких. (Интересно, насколько личным был спор о вреде курения для участвовавших в нем ученых. Фишер так никогда и не расстался со своей трубкой, а Ерушалми не отказался от сигарет.)

Корнфилд прямо нацелился на конституциональную гипотезу Фишера и выбрал для этого вотчину самого Фишера — математику. Предположим, спорил он, что имеется некий конфаундер, например ген курильщика, который полностью

отвечает за риск возникновения рака у курильщиков. Если у курящих риск заболеть раком легких в девять раз выше, фактор-осложнитель должен по крайней мере в девять раз чаще встречаться именно у курильщиков, чтобы разницу в риске возможно было им объяснить. Подумаем, что это значит. Если ген курильщика есть у 11% некурящих, тогда он должен быть у 99% курильщиков. Если ген рака легких встречается хотя бы у 12% некурящих, тогда становится математически невозможно, чтобы исключительно ген рака легких определял ассоциацию между заболеванием и курением. Для биологов этот аргумент, названный неравенством Корнфилда, превратил конституциональную гипотезу Фишера в дымящиеся руины. Нельзя было себе представить, чтобы генетическая изменчивость была так жестко привязана к такому сложному и непредсказуемому предмету, как выбор человека, курить ему или нет.

Неравенство Корнфилда было на самом деле каузальным аргументом, хоть и в зачаточном состоянии: оно предоставляет нам критерий для выбора между диаграммой на рис. 29 (в которой конституциональная гипотеза не способна полностью объяснить ассоциацию между курением и раком) и диаграммой на рис. 30 (на которой ген курильщика полностью определяет наблюдаемую связь).

Рис. 29. Диаграмма, отражающая влияние гена курильщика на привычку курить и заболевание раком

Рис. 30. Диаграмма, отражающая полное влияние гена курильщика на привычку курить и заболевание раком

Как объяснялось выше, связь между курением и раком легких слишком сильна для того, чтобы ее можно было объяснить конституциональной гипотезой.

На самом деле метод Корнфилда заложил основу для очень мощной техники, называемой анализом чувствительности, которая сегодня поддерживает выводы, полученные благодаря машине вывода, описанной во введении. Вместо того чтобы делать умозаключения, предполагая отсутствие определенных каузальных взаимоотношений в модели, аналитик проверяет эти допущения и оценивает, насколько прочными должны быть альтернативные связи, чтобы ими объяснять наблюдаемые данные. Количественный результат затем подвергается оценке на правдоподобие, примерно как предварительные грубые суждения, использованные для обоснования отсутствия этих причинных связей. Нужно ли говорить, что если мы захотим распространить подход Корнфилда на модель с числом переменных больше трех или четырех, нам понадобятся алгоритмы и методы оценки, немыслимые без применения современных графических инструментов.

Эпидемиологов в 50-х годах XX века постоянно критиковали за то, что их доказательства — это «голая статистика». Недоставало якобы лабораторных экспериментов. Но даже беглый взгляд на историю показывает, что этот аргумент лицемерен. Если бы стандарты клинических испытаний применили к цинге, моряки продолжали бы умирать вплоть до 30-х годов XX века, потому что до открытия витамина С не существовало клинических доказательств, что плоды цитрусовых предотвращают цингу. Более того, в 50-х годах XX века некоторые лабораторные подтверждения влияния курения на развитие рака начали появляться в медицинских журналах. У крыс, которых мазали сигаретной смолой, возникали опухоли. Было показано, что сигаретный дым содержит бензопирены — вещества, о канцерогенности которых было известно и ранее. Эти эксперименты подтвердили биологическое правдоподобие гипотезы, что курение может провоцировать онкологию.

К концу десятилетия накопившиеся доказательства самых разных типов убедили почти всех экспертов в этой области, что

курение действительно вызывает рак. Что характерно, убедить удалось даже исследователей из табачных компаний — факт, который оставался глубоко под ковром до 90-х годов XX века, когда скандальные судебные процессы и выступления обличителей вынудили производителей табака рассекретить тысячи прежде закрытых документов. В 1953 году, например, химик Клод Тигль, работавший на «Р.Дж. Рейнольдс тобако компани», написал руководству компании, что табак «является важным этиологическим фактором в индукции первичного рака легкого», что почти буквально совпадало с выводами Долла и Хилла.

На публику, однако, производители сигарет пели совсем другие песни. В январе 1954 года ведущие табачные компании (включая «Рейнольдс») опубликовали в газетах всей страны рекламу «Честное заявление курильщикам сигарет», которое гласило: «Мы глубоко убеждены, что наша продукция не вредит здоровью. Мы всегда сотрудничали и будем сотрудничать с теми, кто стоит на страже здоровья нашего народа». В своей речи в марте 1954 года Джордж Вейссман, президент «Филип Моррис и Ко», заявил: «Если бы у нас появилась информация или хотя бы догадка о том, что продаваемая нами продукция вредит потребителям, мы закрылись бы уже на следующий день». Прошло уже больше 60 лет, а мы все ждем, когда же компания «Филип Моррис» выполнит свое обещание.

Это подводит нас к самому грустному эпизоду во всей истории про курение и рак: сознательным попыткам табачных компаний обмануть людей, скрыв от них риски для здоровья. Если природа подобна джинну, который всегда отвечает на вопросы честно, но исключительно в соответствии с буквой вопроса, представьте, насколько сложнее ученым оказаться лицом к лицу с противником, который сознательно хочет нас обмануть. Сигаретные войны были первым столкновением науки с организованным отрицанием, и никто не был к этому готов. Табачные компании раздували до небес любые противоречащие научные данные, которые им удавалось найти. Они основали собственный Исследовательский комитет табачной индустрии — организацию, которая выделяла ученым деньги на проекты по изучению вопросов, связанных

с табаком и раком, но каким-то удивительным образом никогда не подбиралась к главному вопросу. Когда им удавалось найти легитимных скептиков, отрицающих связь рака и курения, таких как Р. Э. Фишер и Якоб Ерушалми, табачные компании платили им за консультации.

С Фишером все было особенно печально. Конечно, скептицизму всегда есть место. Статистикам платят за то, чтобы они были скептиками — они представляют собой сознание науки. Однако есть разница между оправданным и неоправданным скептицизмом. Фишер пересек эту невидимую линию и прошел дальше. Не способный признать собственные ошибки и определенно находящийся под воздействием привычки к курению, сопровождавшей его всю жизнь, он не мог признать, что вал доказательств обернулся против него. Его возражения приобретали все более отчаянный характер. Он до последнего цеплялся за единственный контринтуитивный результат из первой статьи Долла и Хилла — наблюдение (которое с трудом дотягивало до уровня статистической значимости), что пациенты с раком легких в среднем сообщали, что не затягиваются глубоко, в отличие от здоровых курильщиков. Ни одно из последующих исследований такого эффекта не обнаружило. Хотя Фишер не хуже других знал, что статистически значимые результаты иногда не получается повторить, он скатился в паясничество. Он стал утверждать, что глубокое вдыхание сигаретного дыма оказывает положительный эффект, и призывать к дальнейшему исследованию этого «очень важного момента». Из хорошего о роли Фишера в табачном споре мы, пожалуй, сможем назвать только то, что он, скорее всего, не продался табачным магнатам — его собственного упрямства было более чем достаточно.

По всем этим причинам отношение к связи между курением и раком оставалось противоречивым для публики еще долго после того, как эпидемиологи пришли по его поводу к согласию. Даже врачи, которым, по идее, следует быть ближе к науке, продолжали сомневаться: опрос, проведенный Американским обществом борьбы с раком в 1960 году, показал, что всего лишь треть докторов страны были согласны с утверждением,

что курение является основной причиной рака легких, а 43% из них сами курили.

Хотя мы имеем полное право порицать Фишера за его закоснелость, а табачные компании — за сознательный обман, мы должны также признать, что научное сообщество тогда функционировало в смирительной рубашке. Фишер был прав, пропагандируя рандомизированные контролируемые эксперименты как высокоэффективной способ выяснения причинно-следственных связей. Однако он и его последователи не смогли осознать, что мы в состоянии многое узнать также и из чисто наблюдательных исследований. В этом преимущество каузальной модели: она мобилизует научные познания экспериментатора. Методы Фишера предполагают, что экспериментатор начинает работу без предварительного знания или мнения по поводу проверяемой гипотезы. Они навязывают ученому невежество, и этой ситуацией с радостью воспользовались отрицатели.

Поскольку у ученых не было прямого определения понятия «причина» и способа подтвердить каузальное воздействие без РКИ, они оказались не готовы к спору о том, вызывает ли курение рак. Им пришлось вслепую пробиваться к этому определению в течение долгого периода, который протянулся через все 1950-е и драматически завершился только в 1964 году.

Комиссия начальника здравоохранения и критерии Хилла

Статья Корнфилда и Лилиенфельда открыла дорогу для уверенных заявлений о вреде курения со стороны органов власти, занимающихся здоровьем. Королевский колледж врачей в Великобритании возглавил эту волну, издав в 1962 году отчет с выводом о том, что курение является причиной рака легких. Вскоре после этого министр здравоохранения Соединенных Штатов Америки Лютер Терри (вполне вероятно, что по поручению президента Джона Ф. Кеннеди) объявил о намерении

создать специальную экспертную комиссию по изучению этого вопроса.

Комиссия была тщательно уравновешена, в нее входили пятеро курящих и пятеро некурящих членов, два человека от табачной индустрии и никого из тех, кто ранее участвовал в дебатах о курении, как с той, так и с другой стороны. Поэтому такие люди, как Лилиенфельд и Корнфилд, туда не вошли. Члены комиссии были признанными экспертами в области медицины, химии или биологии. Один из них, Уильям Кохран из Гарвардского университета, был статистиком. На самом деле репутация Кохрана в статистике была лучшей из возможных: он был учеником ученика Карла Пирсона.

Комиссия работала над отчетом больше года, и одной из самых больших проблем для нее оказалось слово «причина». Членам комиссии пришлось отказаться от детерминистских концепций причинности родом из *XIX* века, а также отложить в сторону статистику. Как они (вероятно, Кохран) писали в отчете, «статистические методы не могут доказать существование причинно-следственных отношений в ассоциации. Каузальное значение ассоциации — это вопрос суждения, которое выходит за рамки любой оценки статистической вероятности. Чтобы судить о каузальной значимости ассоциации между признаком, или агентом, и заболеванием, или воздействием на здоровье, нужно использовать целый ряд критериев, ни один из которых не является необходимым и достаточным основанием для такого суждения». Комиссия использовала пять таких критериев: согласованность (многие исследования в различных популяциях показывают похожие результаты); сила ассоциации (включая эффект «доза — ответ»: большее потребление табака ведет к большему риску); специфичность ассоциации (конкретный агент должен производить конкретный эффект, а не длинный список разных эффектов); временные отношения (эффект должен следовать за причиной); непротиворечивость (биологическая правдоподобность и согласованность с другими типами доказательств, такими как лабораторные эксперименты и временные ряды).

В 1965 году Остин Бредфорд Хилл, не состоявший в комиссии, попытался просуммировать эти аргументы таким образом, чтобы их можно было применять к другим проблемам общественного здоровья, и добавил к списку еще четыре критерия: в результате весь список из девяти критериев стал известен как «критерии Хилла». На самом деле Хилл называл их позициями, не требованиями, и подчеркивал, что каждая из них может отсутствовать в конкретном случае: «Ни одна из девяти моих позиций не может дать неопровержимого доказательства или опровержения гипотезы причинно-следственной связи, и ни одна не является обязательным условием», — писал он.

Действительно, несложно найти аргументы против каждого из критериев как из списка Хилла, так и из более короткого списка экспертной комиссии. Согласованность сама по себе ничего не доказывает: если 30 исследований игнорируют один и тот же конфаундер, все они легко могут оказаться со смещенной оценкой. Сила ассоциации уязвима по той же причине; как упоминалось ранее, размер обуви детей сильно коррелирует с их умением читать, но не связан с ним каузально. Специфичность всегда была особенно противоречивым критерием. Она имеет четкий смысл в контексте инфекционных болезней, когда один возбудитель, как правило, вызывает одно заболевание, но уже сильно размыта в случае средовых влияний. Курение повышает риск целого ряда других недугов, таких как эмфизема и сердечно-сосудистые заболевания. Снижает ли это надежность данных, что оно вызывает рак? Временная связь тоже не лишена исключений, например, рассвет наступает не из-за пения петуха, хотя петух поет всегда перед рассветом.

Наконец, согласованность с известными теориями и фактами, конечно, желательна, но в истории науки масса опровергнутых теорий и ошибочных лабораторных открытий.

Положения Хилла все еще полезны в качестве описания того, как научная дисциплина подходит к принятию каузальной гипотезы, используя различные типы подтверждений, но методология, которая позволила бы их применять, отсутствует. Так, биологическое правдоподобие и согласованность с экспериментальными данными, вероятно, хорошие, нужные

вещи. Но как именно нам следует определять вес подобных подтверждений? Как именно мы вставим имевшиеся ранее знания в новую картину? Очевидно, на эти вопросы каждый ученый должен отвечать самостоятельно. Однако интуитивные решения могут быть ошибочными, особенно если в игру вступают политическое давление, финансовые выгоды или же если исследователь находится в зависимости от вещества, которое изучает.

Конечно, ни один из этих комментариев не предполагал хоть в чем-то принизить работу комиссии. Ее состав сделал все возможное в условиях отсутствия механизмов для обсуждения причинности. Их вывод о том, что нужны и нестатистические критерии, был огромным шагом вперед, а сложные личные решения, принятые курящими членами комитета, подтвердили серьезность их заключений. Лютер Терри, куривший сигареты, перешел на трубку. Леонард Шуман объявил, что бросает курить. Уильям Кохран признал, что, бросив курить, снизил бы риск заболеть раком, но чувствует, что ощущение комфорта, которое доставляют сигареты, оправдывает риск. Печальнее всего было то, что у Луиса Физера, выкуривавшего четыре пачки в день, обнаружили рак легких меньше чем через год после отчета. Он писал комиссии: «Вы, вероятно, помните, что, хотя меня и полностью убедили приведенные доказательства, я продолжал много курить во время всей работы комиссии, находя обычные оправдания… Мой собственный случай кажется мне убедительней любой статистики». Потеряв одно легкое, он все-таки бросил курить.

С позиций общественного здравоохранения отчет экспертной комиссии был эпохальным явлением. В течение двух лет конгресс ввел требование к производителям табака поместить на сигаретные пачки предупреждения об угрозе здоровью. В 1971 году рекламу сигарет запретили на радио и телевидении. Процент курящих среди взрослого населения США снизился с исторического максимума в 45,0% в 1965 году до 19,3% в 2010 году. Кампания против курения была одним из крупнейших и самым успешным, пусть и мучительно медленным и незавершенным достижением общественного здра-

воохранения в истории. Комиссия также выработала ценный протокол для достижения научного консенсуса и послужила моделью для будущих отчетов начальника здравоохранения по теме курения и многим другим в последующие годы (включая пассивное курение, которое стало большой проблемой в 80-х годах XX века).

С точки зрения причинности этот отчет был в лучшем случае весьма скромным успехом. Он ясно обозначил серьезность каузальных вопросов и то, что голые данные неспособны на них ответить. Но в качестве дорожной карты для будущих открытий его руководящие принципы не вполне годились из-за их неопределенности и неуклюжести. Критерии Хилла лучше всего воспринимать как исторический документ, суммирующий типы доказательств, возникшие в 50-х годах XX века и сумевшие убедить медицинское сообщество. Но они не годятся как руководство для будущих исследований. Для любых каузальных вопросов, кроме разве что самых общих, нам нужен более точный инструмент. Оглядываясь назад, неравенство Корнфилда, которое посеяло семена анализа сенситивности, было шагом в этом направлении.

Курение для новорожденных

Даже после того, как горячие споры по поводу курения и рака улеглись, один крупный парадокс продолжал будоражить умы. Якоб Ерушалми указал на то, что курение матери во время беременности, по всей видимости, шло на пользу здоровью ребенка, если он родился с недостаточным весом. Эта загадка, известная как парадокс веса при рождении, была плевком в лицо нарождающемуся медицинскому консенсусу относительно курения, и его не удавалось удовлетворительно объяснить вплоть до 2006 года — спустя более 40 лет после выхода публикации Ерушалми. Я абсолютно уверен, что это заняло столько времени потому, что язык причинности был недоступен с 1960 по 1990 год.

В 1959 году Ерушалми начал долговременное исследование общественного здоровья, которое собрало пре- и постнатальные данные о более чем 15 тысячах детей в районе залива Сан-Франциско. Эти данные включали также информацию о том, курили ли матери этих детей, а также вес и смертность младенцев в течение первого месяца жизни.

Несколько предыдущих работ уже показали, что дети курящих матерей при рождении весят в среднем меньше, чем дети некурящих, и было бы естественно предположить, что у них и выживаемость должна быть хуже. В самом деле, исследование детей с недостаточным весом (определяемым как менее 5,5 фунтов при рождении) по всей стране показало, что их смертность более чем в 20 раз выше, чем у детей с нормальным весом. Таким образом, эпидемиологи выстроили цепочку причинно-следственных связей: курение → низкий вес при рождении → смертность.

То, что обнаружил Ерушалми, обрабатывая данные, оказалось сюрпризом даже для него самого. Дети курящих матерей действительно были в среднем легче, чем дети некурящих (примерно на 7 унций). Однако дети с недостаточным весом, родившиеся у курящих матерей, выживали лучше, чем дети некурящих из этой же категории. Выглядело это так, будто курение матери на самом деле обладало защитным воздействием.

Если бы нечто подобное нашел Фишер, он бы немедленно во всеуслышание объявил это пользой от курения. Ерушалми, надо отдать ему должное, так себя не повел. Он написал, очень осторожно: «Это парадоксальное открытие вызывает сомнения и противоречит предположению, что курение действует как экзогенный фактор, который взаимодействует с внутритробным развитием плода». Говоря короче, от переменной *курение* к переменной *смертность* нет каузального пути.

Современные эпидемиологи полагают, что Ерушалми был неправ. Большинство из них полагает, что курение все-таки увеличивает смертность новорожденных, например, потому, что взаимодействует с переносом кислорода через плаценту. Но как нам примирить эту гипотезу с такими данными?

Статистики и эпидемиологи настаивали на анализе этого парадокса в терминах вероятности и восприятии его как ано-

малии, свойственной весу при рождении. Оказалось, что это явление имеет слабое отношение к натальной массе, зато четко связано с коллайдерами. Если рассматривать его в этом свете, оно вовсе не парадоксально, а показательно.

На самом деле данные Ерушалми прекрасно согласуются с моделью «курение → низкий вес при рождении → смертность», если добавить к ней кое-что еще. Курение действительно причиняет вред, являясь причиной низкого веса при рождении, однако некоторые другие причины этого, такие как серьезные или угрожающие жизни генетические аномалии, приносят гораздо больше вреда. Низкому весу при рождении у данного конкретного ребенка есть два объяснения: его мать могла курить или же повлияла одна из этих прочих причин. Если мы узнаем, что его мать курила, эта информация вполне объясняет низкий вес и, следовательно, снижает вероятность серьезного нарушения развития. Но, если мать не курила, у нас есть гораздо более серьезное свидетельство в пользу того, что причина низкого веса — это нарушения развития, и дальнейший прогноз для ребенка становится мрачнее.

Как и раньше, с каузальной диаграммой все становится понятнее. Когда мы подключаем новые допущения, она начинает выглядеть как на рис. 31. Мы видим, что парадокс веса при рождении — это прекрасный пример ошибки оценки, возникающей при коллайдере. Переменная, к которой сходятся пути, — это, собственно, вес при рождении. Если мы берем только детей с низкой натальной массой, мы вводим поправку по этой переменной. Это открывает черный ход между *курением* и *смертностью* по схеме «курение → низкий вес при рождении ← нарушения внутриутробного развития → смертность». Этот путь — некаузальный, потому что одна из стрелок направлена не в ту сторону. Однако он вызывает ложную корреляцию между *курением* и *смертностью* и смещает нашу оценку реальной (прямой) причинно-следственной связи «курение → смертность». На самом деле, здесь он смещает оценку так сильно, что курение даже кажется благотворным.

Красота каузальных диаграмм в том, что они делают источник ошибки очевидным. Без этих диаграмм эпидемиологи спорили об этом парадоксе 40 лет. На самом деле этот случай

обсуждается до сих пор: в «Международном журнале эпидемиологии» за октябрь 2014 года содержится сразу несколько статей на эту тему. Одна из них, написанная Тайлером ВандерВиле из Гарварда, четко расписывает объяснение парадокса и приводит точно такую же диаграмму, как та, что приведена ниже.

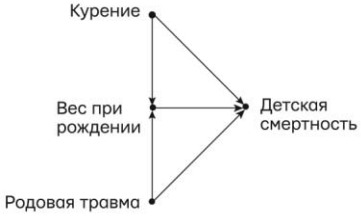


Рис. 31. Каузальная диаграмма для парадокса веса при рождении

Конечно, эта диаграмма слишком упрощена и не описывает все детали взаимосвязей курения, веса при рождении и младенческой смертности. Тем не менее принцип смещения оценки при коллайдере очень действенен. В этом случае смещение было обнаружено потому, что наблюдаемое явление было слишком уж неправдоподобно, но только представьте себе, насколько часто смещение при коллайдере остается незамеченным, потому что не противоречит теории.

Страстные дебаты: наука против культуры

Уже после того, как я начал работать над этой главой, мне предоставился случай связаться с Алленом Вилкоксом, эпидемиологом, чье имя, пожалуй, вспоминается первым в связи с этим парадоксом. Он задал очень неудобный вопрос про диаграмму на рис. 31: откуда мы знаем, что низкий вес при рождении — это действительно прямая причина смертности? Он считает, что на самом деле врачи всегда интерпретировали малую натальную массу неверно. Поскольку он сильно связан с младенческой смертностью, врачи полагают его причиной.

В действительности эта связь может полностью обусловливаться конфаундерами (представленными как «нарушения развития» на рис. 31, хотя Вилкокс рассматривает вопрос как более общий).

По поводу аргумента Вилкокса имеет смысл сделать два замечания. Во-первых, даже если мы удаляем стрелку *вес при рождении → смертность*, коллайдер остается. Таким образом, каузальная диаграмма продолжает успешно объяснять парадокс веса при рождении. Во-вторых, каузальная переменная, на которой были сосредоточены исследования Вилкокса, — не курение, а раса. А расовые вопросы по-прежнему вызывают страстные споры в нашем обществе.

Оказывается, тот же самый парадокс веса при рождении, что у детей курильщиц, наблюдается и у чернокожих матерей. У последних дети с недостаточным весом рождаются чаще, чем у белых, и в целом у их детей выше младенческая смертность. Однако их дети с недостаточным весом выживают лучше, чем белые дети с недостаточным весом. Какой же вывод мы должны из этого сделать? Мы можем рекомендовать курящей беременной женщине перестать курить, чтобы не вредить ребенку, но немыслимо советовать беременной женщине сменить расовую принадлежность.

Вместо этого нам следует обратить внимание на социальные проблемы, из-за которых у детей черных матерей смертность оказывается выше. Это утверждение непротиворечиво. Но какие именно причины в данном случае важнее всего и в чем следует измерять успех? Хорошо это или плохо, но многие борцы за расовую справедливость полагают вес при рождении промежуточным звеном в каузальной цепочке «раса → вес при рождении → смертность». Более того, они берут натальную массу в качестве замещающей переменной вместо младенческой смертности, предполагая, что если к лучшему изменится одно, то автоматически к лучшему изменится и второе. Понять, почему возникла такая практика, несложно: данные по весу при рождении гораздо более доступны, чем данные по младенческой смертности.

Теперь представим, что произойдет, если кто-нибудь, подобно Вилкоксу, придет и заявит, что низкий вес при рождении

сам по себе не является медицинским состоянием и не имеет причинностной связи с младенческой смертностью. Это развалит весь терем-теремок. Вилкокса обвинили в расизме еще тогда, когда он впервые предложил эту идею, в 1970-х, и он не осмеливался опубликовать ее до 2001 года. И даже тогда статья вышла в сопровождении двух комментариев, и один из них поднимал расовый вопрос: «В контексте общества, где доминирующая группа оправдывает свое господство, доказывая генетическую ущербность тех групп, над которыми она господствует, сложно держать нейтралитет, — написал Ричард Дэвид из госпиталя Кук Каунти в Чикаго. — В погоне за „чистой наукой" исследователь с самыми лучшими намерениями может восприниматься — а иногда и быть — тем, кто поддерживает и укрепляет ненавистное ему социальное устройство».

Это жесткое обвинение, вырастающее из самых благородных побуждений, — не первый случай, когда ученому доставалось за выяснение истины, потенциально несущей неблагоприятные последствия для общества. Возражения Ватикана на идеи Галилея очевидно были вызваны искренними опасениями за общественный порядок того времени. То же самое можно сказать и об идее эволюции Дарвина, и о евгенике Фрэнсиса Гальтона. Однако культурный шок, возникающий вследствие научных открытий, сглаживается соответствующими изменениями в культуре, постепенно встраивающей эти открытия в себя, а вовсе не запретами. Важное, необходимое условие для подобной подстройки — умение отделить науку от культуры до того, как вспыхнет война мнений. К счастью, язык каузальных диаграмм сегодня предоставляет нам возможность бесстрастно рассуждать о причинах и следствиях не только тогда, когда все лежит на поверхности, но и в самых сложных случаях.

«Парадокс Монти Холла», трудная и многих даже приводящая в бешенство задача, которая демонстрирует, как наш ум легко вводится в заблуждение вероятностными рассуждениями там, где на самом деле следует пользоваться логикой причинности.

Глава 6
Сплошные парадоксы!

Тот, кто сталкивается с парадоксальным,
оказывается лицом к лицу с реальностью.
Фридрих Дюрренматт, 1962

Парадокс веса при рождении, на котором мы закончили главу 5, относится к удивительно объемному классу парадоксов, отражающих противоречия между причинно-следственными связями и ассоциациями. Напряжение возникает, поскольку они находятся на двух разных уровнях Лестницы Причинности, и усугубляется, потому что человеческая интуиция следует логике причинно-следственных связей, в то время как данные повинуются логике вероятностей и соотношений. Парадоксы появляются, когда мы неверно применяем правила, усвоенные в одной области, к другой.

Мы посвятим эту главу очень известным и загадочным парадоксам, связанным с вероятностью и статистикой, прежде всего, потому что это интересно. Если вы не знакомы с парадоксами Монти Холла и Симпсона, могу обещать хорошую тренировку для мозга. И даже если вы считаете, что знаете о парадоксах все, думаю, будет интересно взглянуть на них сквозь призму причинности, благодаря которой все выглядит иначе.

В то же время мы изучаем парадоксы не только потому, что они забавляют и развлекают. Подобно оптическим иллюзиям,

они показывают, как работает наш мозг, какие уловки он использует и что создает для него проблемы.

Причинные парадоксы проливают свет на шаблоны интуитивных рассуждений о причинно-следственных связях, которые вступают в противоречие с логикой вероятности и статистики. До такой степени, что приходится нелегко даже статистикам. Мы увидим, как порой они садились в калошу, и это послужит нам предупреждением: взгляд на мир без каузальной оптики может быть ошибочным.

Заковыристая задача Монти Холла

В конце 1980-х журналистка по имени Мэрилин вос Савант начала вести регулярную колонку в журнале «Парад», еженедельном приложении к воскресным газетам во многих городах США. Колонка «Спросите Мэрилин» выходит по сей день, и в ней даются ответы на загадки, головоломки и научные вопросы, предложенные читателями. Журнал объявил вос Савант «самой умной женщиной в мире», что, несомненно, мотивирует читателей искать вопросы, которые поставили бы ее в тупик.

За время существования колонки больше всего шуму наделала задача, опубликованная в сентябре 1990 года: «Представьте, что вы участвуете в телевизионной игре и вам на выбор предлагают три двери. За одной дверью стоит автомобиль, за двумя другими — козы. Допустим, вы выбираете дверь 1, и ведущий, который знает, что где спрятано, открывает дверь 3. За ней коза. Потом ведущий спрашивает, не хотите ли вы теперь выбрать дверь 2. Стоит ли менять выбор?»

Американским читателям было очевидно, что задача отсылает к популярной телеигре под названием «Заключим сделку» (Let's Make a Deal), ведущий которой, Монти Холл, предлагал участникам именно такие вопросы. Вос Савант решила, что выбор стоит изменить: если не менять дверь, шанс на выигрыш будет равен одному из трех, а если поменять, он удвоится и будет равен двум из трех.

Даже умнейшая женщина в мире не могла предвидеть дальнейшее развитие событий. За следующие месяцы она получила более 10 тысяч писем от читателей — в основном не согласных с ее решением. Немало ответов пришло от людей, которые утверждали, что имеют докторскую степень по математике или статистике. И вот что писали ученые: «Это провал, настоящий провал!» (доктор Скотт Смит); «Я бы предложил вам взять стандартный учебник по теории вероятности и почитать его, прежде чем вы снова возьметесь за вопросы такого типа» (доктор Чарльз Рид); «Это провал!» (доктор Роберт Сакс) и «Вы совершенно неправы» (доктор Рэй Бобо). В общем и целом критики утверждали, что не было никакой разницы, поменяет ли участник решение, — в игре осталось две двери, первоначальный выбор был абсолютно случайным, и вероятность, что машина окажется за любой из дверей, неизменно равна одной второй.

Кто был прав? А кто неправ? И почему задача вызвала такие страсти? Все три вопроса заслуживают внимательного рассмотрения. Давайте сначала посмотрим, как справилась с задачей вос Савант. Ее решение поражает простотой и выглядит убедительнее, чем все, что я видел в многочисленных учебниках. Она составила список из трех возможных вариантов для расположения дверей и коз, а также внесла соответствующие результаты для стратегий «Поменять» и «Не менять» (табл. 5). Во всех трех случаях предполагается, что сначала вы выбрали дверь 1. Поскольку все перечисленные возможности (изначально) равновероятны, шанс на выигрыш, если сменить дверь, составляет две трети, а если не менять — одну треть. Обратите внимание, что в таблице вос Савант явно не указано, какую дверь открыл ведущий. Эта информация подразумевается в колонках 4 и 5. Например, во второй строке мы учли, что ведущий должен открыть дверь 3, поэтому смена выбора приведет вас к двери 2 и выигрышу. Точно так же в первой строке открытой может быть дверь 1 или дверь 2, но в столбце 4 правильно указано, что, сменив выбор, вы проиграете в любом случае. Даже сегодня многие люди, которые впервые видят эту

головоломку, не могут поверить в результат. Почему? Какой интуитивный нерв им защемили?

Таблица 5. Три сочетания дверей и коз в «Заключим сделку» показывают, что вариант сменить дверь вдвое привлекательнее

Дверь 1	Дверь 2	Дверь 3	Результат, если поменять дверь	Результат, если не менять дверь
Автомобиль	Коза	Коза	Проигрыш	Выигрыш
Коза	Автомобиль	Коза	Выигрыш	Проигрыш
Коза	Коза	Автомобиль	Выигрыш	Проигрыш

Вероятно, на это есть 10 тысяч разных причин, по одной на каждого читателя, но я считаю самым убедительным аргументом такой: нам кажется, что решение вос Савант принуждает верить в телепатию. Если я должен изменить выбор независимо от того, какую дверь предпочел в начале, значит, продюсеры каким-то образом читают мои мысли. А иначе как им удастся расположить машину таким образом, чтобы она с большей вероятностью оказалась за дверью, которую я не выбрал?

Ключевой элемент для разрешения этого парадокса состоит в том, что необходимо принять в расчет не только данные (т.е. факт, что ведущий открыл конкретную дверь), но и процесс генерации данных, другими словами, правила игры. Они сообщают кое-что об информации, которую можно было бы получить, только мы этого не сделали. Неудивительно, что именно статистикам было так трудно понять решение головоломки. Они привыкли к тому, что Р. Э. Фишер в 1922 году

назвал сокращением данных, и традиционно игнорируют процесс их генерации.

Для начала давайте немного поменяем правила игры и посмотрим, как это повлияет на наш выбор. Представьте альтернативную игру с ведущим по имени Хонти Молл. Он открывает одну из дверей, на которую вы не показывали, но его выбор абсолютно случаен, т.е. он может открыть и ту дверь, за которой находится автомобиль. Такая вот незадача!

Чтобы рассмотреть этот вариант, составим таблицу, похожую на предыдущую, и примем во внимание факт, что расположение машины (три возможности) и выбор двери, которые сделает Хонти Молл (две возможности) — два случайных и независимых обстоятельства. Таким образом, в таблице должно быть шесть строк, каждая из которых равновероятна, потому что обстоятельства независимы друг от друга (табл. 6).

Что же произойдет, если Хонти Молл откроет дверь и за ней окажется коза? Так у нас появится важная информация: вероятно, мы находимся на второй или четвертой строке таблицы. Сосредоточившись только на второй и четвертой строках, мы увидим, что отказ от изначального выбора больше не дает нам никакого преимущества — в любом случае вероятность выигрыша будет равна одной второй. Таким образом, применительно к шоу Хонти Молла критики Мэрилин вос Савант были бы правы! Однако в двух этих случаях мы имеем дело с разными данными. Урок здесь довольно прост: то, как мы получаем информацию, не менее важно, чем сама информация.

Таблица 6. Варианты в программе с Хонти Моллом

Ваш выбор	Дверь с автомобилем	Дверь, открытая ведущим	Результат, если поменять дверь	Результат, если не менять дверь
1	1	2 (коза)	Проигрыш	Выигрыш
1	1	3 (коза)	Проигрыш	Выигрыш

Окончание таблицы

Ваш выбор	Дверь с автомобилем	Дверь, открытая ведущим	Результат, если поменять дверь	Результат, если не менять дверь
1	2	2 (автомобиль)	Проигрыш	Проигрыш
1	2	3 (коза)	Выигрыш	Проигрыш
1	3	2 (коза)	Выигрыш	Проигрыш
1	3	3 (автомобиль)	Проигрыш	Проигрыш

Давайте воспользуемся нашим любимым приемом и нарисуем диаграмму причинности, которая сразу же покажет, чем отличаются две телеигры, реальная и воображаемая. Во-первых, на рис. 32 показана диаграмма для настоящей игры «Заключим сделку», где Монти Холл должен открыть дверь, за которой нет машины. Отсутствие стрелки между вашей дверью и местонахождением автомобиля означает, что ваш выбор двери и выбор продюсера, где поставить машину, не зависят друг от друга. Так мы явно исключаем возможность чтения мыслей со стороны продюсеров (или с вашей стороны!). Но еще важнее две стрелки, присутствующие на диаграмме. Они показывают, что на открывшуюся дверь повлиял и ваш выбор, и выбор продюсеров. Дело в том, что Монти Холл должен открыть дверь, отличную и от «вашей двери», и от «местонахождения автомобиля» — ему необходимо принять в расчет оба фактора.

Рис. 32. Диаграмма причинности для шоу Монти Холла

Как следует из рис. 32, открытая дверь — это коллайдер. Как только у нас появляются данные об этой переменной, все наши вероятности становятся зависимыми от полученной информации. Но, если учитывать только коллайдер, создается ложная зависимость между его родителями. Зависимость подкрепляется вероятностями: если вы выбрали дверь 1, то вероятность присутствия автомобиля за дверью 2 будет в два раза выше, чем шанс найти его за дверью 1; если вы выбрали дверь 2, то в два раза выше вероятность его присутствия за дверью 1.

Это определенно странная зависимость, к которой большинство из нас не привыкло. У этой зависимости нет причины. Она не связана с физическим общением между продюсерами и нами. Она не подразумевает телепатию. Это чистый эффект байесовской обусловленности: волшебная передача информации без причинности. Наш разум восстает против такой возможности, потому что с раннего детства мы научились связывать корреляцию с причинностью. Если машина позади нас делает те же повороты, что и мы, сначала мы думаем, что она следует за нами (причинно-следственная связь!). Потом нам приходит в голову, что она просто едет в то же место (т.е. за каждым поворотом стоит общая причина). Но беспричинная корреляция противоречит здравому смыслу. Таким образом, парадокс Монти Холла подобен оптической иллюзии или фокусу: он использует наши собственные когнитивные механизмы, чтобы обмануть нас.

Почему я говорю, что, когда Монти Холл открыл дверь 3, произошла «передача информации»? Ведь вы не получили никаких подтверждений того, что поступили верно, выбрав дверь 1. Вы заранее знали, что он собирается открыть дверь, за которой спрятана коза, и это случилось. Никто не попросит вас изменить это мнение, поскольку вы стали свидетелем неизбежного. Так почему же вероятность выигрыша за дверью 2 выросла с $\frac{1}{3}$ до $\frac{2}{3}$?

Дело в том, что Монти не мог открыть дверь 1 после того, как вы ее выбрали, но мог открыть дверь 2. Поскольку этого не произошло, вероятность, что он открыл дверь 3 по необхо-

димости, повышается. Таким образом, мы получаем больше подтверждений того, что автомобиль находится за дверью 2. Это общая тема байесовского анализа: любая гипотеза, выдержавшая какую-то проверку на достоверность, становится вероятнее. Чем больше угроза достоверности, тем выше вероятность после ее преодоления. Вероятность двери 2 могла быть опровергнута (т.е. Монти мог ее открыть), а двери 1 — нет. Таким образом, дверь 2 становится более вероятным местом. Вероятность того, что машина находится за дверью 1, остается $\frac{1}{3}$.

Для сравнения на рис. 33 представлена диаграмма причинности для альтернативной игры с ведущим Хонти Моллом, который выбирает не ту дверь, что вы, но в остальном делает это наугад. На этой диаграмме все еще есть стрелка, указывающая от вашей двери к открытой двери, потому что он не может выбрать ту же самую. Но стрелка от местоположения машины к открытой двери будет удалена, потому что ведущему неважно, где находится машина. На этой диаграмме открытая дверь никак не влияет на ситуацию: ваша дверь и расположение машины были независимы с самого начала и останутся независимыми, когда мы увидим, что за дверью Хонти. Итак, в телеигре с Хонти Моллом вероятность обнаружить машину за вашей дверью и за другой дверью абсолютно одинакова, что показано в рис. 40.

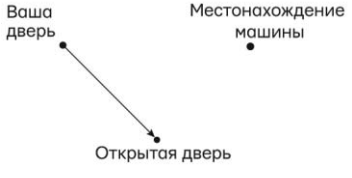


Рис. 33. Диаграмма причинности для шоу Хонти Молла

С байесовской точки зрения разница между этими двумя играми состоит в том, что в шоу Хонти Молла дверь 1 может быть показана как неверный выбор. Монти Холл мог открыть дверь 3 и показать машину, что доказало бы вашу неправоту.

Поскольку ваша дверь и дверь 2 могли быть обозначены как неверные, вероятность для них остается одинаковой.

Хотя это чисто качественный анализ, его можно сделать количественным, использовав правило Байеса или рассматривая диаграммы как простые байесовские сети. Поступая так, мы помещаем эту задачу в общие рамки, которые использовались в подобных случаях. Нам не нужно изобретать метод для решения головоломки; распространение убеждений, описанное в главе 3, даст правильный ответ, а именно P (*дверь* 2) = $\frac{2}{3}$ для шоу Монти Холла и P (*дверь* 2) = ½ для шоу Хонти Молла.

Заметьте, что на деле я предложил два объяснения для парадокса Монти Холла. В первом ложная зависимость между вашей дверью и местонахождением автомобиля объясняется с помощью причинно-следственных связей. Во втором используется байесовский подход, чтобы выяснить, почему в «Заключим сделку» повышается вероятность двери 2. Оба объяснения имеют ценность. Байесовский подход объясняет само явление, но не показывает, почему мы воспринимаем его так парадоксально. На мой взгляд, без ответа на этот вопрос нельзя полностью разрешить парадокс. Почему читатели колонки вос Савант были так уверены, что она ошибалась? Ведь это были не только люди, безосновательно убежденные в своей правоте. Пал Эрдёш, один из выдающихся математиков современности, тоже не мог согласиться с ее решением, пока его не подтвердила компьютерная симуляция. О каких недостатках в нашем интуитивном видении мира говорит эта ситуация?

«Наши мозги не слишком приспособлены для решения задач о вероятности, поэтому ошибки в этой ситуации меня не удивляют», — сказал Перси Диаконис, специалист по статистике из Стэнфордского университета в интервью «Нью-Йорк таймс» в 1991 году. Справедливо, только этим дело не ограничивается. Наши мозги действительно не приспособлены для решения задач о вероятности, однако приспособлены для решения задач о причинно-следственных связях. И эта предрасположенность порождает систематические ошибки в оценке вероятностей — такие же, как в случае с оптическими иллюзиями. Поскольку между «моей дверью» и «расположением автомобиля» нет при-

чинно-следственных связей, нам чрезвычайно трудно понять, что здесь существует вероятностная ассоциация. Наши мозги не готовы принимать беспричинную корреляцию, и, чтобы видеть ситуации, где она возможна, требуется специальная подготовка — на примерах вроде парадокса Монти Холла или тех, что мы обсуждали в главе 3. Как только мы «перезарядим» мозги и начнем узнавать коллайдеры, такие парадоксы перестанут сбивать нас с толку.

И снова об «ошибке коллайдера»: парадокс Берксона

В 1946 году Джозеф Берксон, биостатистик из клиники Мэйо, указал на интересную особенность, которую выявили наблюдения в условиях больницы: даже если два заболевания не связаны друг с другом у населения в целом, они могут показаться связанными у пациентов.

Чтобы понять суть этого наблюдения, давайте начнем с диаграммы причинности (рис. 34). Также полезно подумать об экстремальном варианте: ни болезнь 1, ни болезнь 2 обычно недостаточно серьезны, чтобы привести к госпитализации, но их сочетания достаточно. В этом случае мы ожидаем, что болезнь 1 будет сильно коррелировать с болезнью 2 у госпитализированных людей.

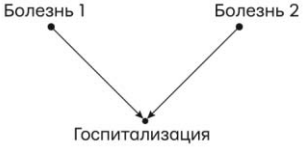


Рис. 34. Диаграмма причинности для парадокса Берксона

Исследуя пациентов в больнице, мы учитываем только госпитализированных людей. Как мы знаем, ограничив себя коллайдером, мы создаем ложную связь между болезнью 1 и болезнью 2. Во многих предыдущих примерах ассоциация была

отрицательной из-за эффекта поверхностного объяснения, но здесь она положительна, потому что для госпитализации требуются оба заболевания (не только одно).

Однако эпидемиологи долгое время отказывались верить в такую возможность. Они игнорировали ее до 1979 года, когда Дэвид Сакетт из Университета Макмастера, эксперт по всевозможным статистическим ошибкам, представил убедительные доказательства того, что парадокс Берксона реален. В одном примере (табл. 7) он изучил две группы заболеваний: респираторные и костные. Около 7,5% людей в общей популяции страдают заболеваниями костей, и этот процент не зависит от того, есть ли у них респираторные заболевания. Но для госпитализированных людей с респираторными заболеваниями частота заболеваний костей возрастает до 25%! Сакетт назвал это явление систематической ошибкой при поступлении в больницу, или систематической ошибкой Берксона.

Таблица 7. Данные Сакетта, иллюстрирующие парадокс Берксона

Наличие респираторного заболевания	Наличие заболевания костей					
	у населения в целом			у госпитализированных за последние шесть месяцев		
	Да	Нет	% Да	Да	Нет	% Нет
Да	17	207	7,6	5	15	25,0
Нет (контрольная группа)	184	2 376	7,2	18	219	7,6

Сакетт признает, что мы не вправе окончательно приписать этот эффект систематической ошибке Берксона, потому что возможны вмешивающиеся факторы. Споры в том или ином виде продолжаются до сих пор. Однако, в отличие от 1946 и 1979 годов, сегодня эпидемиологи понимают причинно-следственные диаграммы и знают, какие систематические ошибки они демонстрируют. Сегодня обсуждаются более тонкие моменты: насколько велика может быть ошибка и достаточно ли она масштабна, чтобы быть замеченной на диаграммах причинности с большим количеством переменных. Это прогресс!

Корреляции, вызванные коллайдером, не новы. Они обнаружились в исследовании, проведенном в 1911 году английским экономистом Артуром Сесилом Пигу, который сравнивал детей, родившихся у алкоголиков и неалкоголиков. Также они встречаются, хотя под другими названиями, в работах Барбары Бёркс (1926), Герберта Саймона (1954) и, конечно же, Берксона. Кроме того, они вовсе не такие эзотерические, как может показаться на моих примерах. Попробуйте такой эксперимент: подбросьте две монеты одновременно 100 раз, но записывайте результаты только тогда, когда хотя бы одна из них выпадет орлом. Посмотрев на таблицу, в который, вероятно, будет около 75 записей, вы обнаружите, что результаты двух одновременных подбрасываний не окажутся независимыми. Каждый раз, когда первая монета выпадала решкой, вторая выпадала орлом. Как это получилось? Неужели монеты общались друг с другом со скоростью света? Конечно нет. На самом деле вы ограничили себя коллайдером, отбросив все результаты «решка — решка».

В книге «Направление времени», опубликованной посмертно в 1956 году, философ Ханс Райхенбах высказал смелую гипотезу, названную принципом общего дела. Опровергая утверждение «Корреляция не подразумевает причинно-следственной связи», Райхенбах выдвинул далеко идущую идею: «Нет корреляции без причинно-следственной связи». Он имел в виду, что корреляция между двумя переменными, X и Y, не может возникнуть случайно. Либо одна из переменных вызывает другую, либо третья переменная, например Z, предшествует и вызывает их обе.

Наш простой эксперимент с подбрасыванием монеты доказывает, что утверждение Райхенбаха пошло слишком далеко, потому что в нем не учитывается процесс отбора наблюдений. У результата для двух монет не было общей причины, и ни одна не сообщала другой, что получилось у нее. Тем не менее между результатами в нашем списке возникла корреляция. Ошибка Райхенбаха заключалась в том, что он не учел структуру коллайдера, на основе которой отбирались данные. Ошибка оказалась особенно показательной, потому что она указывает на конкретный изъян в принципах работы нашего мозга. Мы живем так, как если бы принцип общей причины соблюдался. Когда мы видим такие закономерности, мы ищем причинное объяснение. Более того, мы жаждем объяснений, которые показали бы нам стабильные механизмы за рамками данных. Больше всего нам подходит объяснение через прямую причинность: *X* вызывает *Y*. Когда оно не подходит, обычно нас удовлетворяет общая причина для *X* и *Y*. Коллайдеры слишком призрачны, чтобы удовлетворить эти причинные аппетиты. Мы все еще хотим узнать механизм, как две монеты координируют свое поведение. Ответ вызывает у нас полное разочарование: они вообще не общаются.

Наблюдаемая нами корреляция — иллюзия в чистейшем и буквальнейшем смысле. Или даже заблуждение, т.е. иллюзия, которую мы сами вызвали у себя, выбирая, какие события включить в отобранные данные, а какие проигнорировать. Важно понимать, что мы не всегда осознаем, что сделали этот выбор, и поэтому часто попадаем в ловушку, созданную ошибкой коллайдера. В эксперименте с двумя монетами выбор был осознанным: я просил не записывать результаты с двумя решками. Но во многих случаях мы не осознаем, что делаем выбор, или же выбор делается за нас. В парадоксе Монти Холла ведущий открывает нам дверь. В парадоксе Берксона неосторожный исследователь берет госпитализированных пациентов из соображений удобства, не осознавая, что таким образом искажает результаты исследования.

Искажающая призма коллайдера не менее распространена в повседневной жизни. Джордан Элленберг в книге «Как ни-

когда не ошибаться» спрашивает: вы когда-нибудь замечали, что среди людей, с которыми вы встречались, привлекательные часто оказывались неприятными личностями? Вместо того чтобы строить сложные психосоциальные теории, рассмотрите простое объяснение. Ваш выбор партнеров все это время зависел от двух факторов: их привлекательности и личных качеств. Вы были готовы завязать отношения с неприятным, но привлекательным человеком или с приятным, но непривлекательным, и, конечно, с приятным и привлекательным. Но только не с неприятным и непривлекательным! То же явление мы наблюдали в примере с двумя монетами, когда вы подвергали цензуре результаты «орел или решка». Это явление создает ложную отрицательную корреляцию между привлекательностью и личностью. Но печальная правда заключается в том, что непривлекательные люди бывают неприятными так же часто, как привлекательные, однако вы никогда этого не узнаете, потому что никогда не станете встречаться с плохим и некрасивым человеком.

Парадокс Симпсона

Теперь, когда мы показали, что телепродюсеры не обладают навыками телепатии и монеты не могут общаться друг с другом, мы можем развенчать еще несколько мифов. Давайте начнем с мифа о «плохом / плохом / хорошем» лекарстве.

Представим себе доктора (назовем его «доктор Симпсон»), который сидит в кабинете и читает о многообещающем новом препарате (лекарстве *D*), который вроде бы сокращает риск сердечного приступа. С радостным предвкушением он изучает данные исследователей в Интернете. Однако радость убавляется, когда он смотрит на данные о пациентах-мужчинах и замечает, что, если они принимают препарат, риск получить сердечный приступ, вообще-то, повышается. «Ага, — говорит он, — вероятно, лекарство *D* очень эффективно для женщин».

Но потом он переходит к следующей таблице, и разочарование сменяется изумлением. «Что это? — восклицает доктор

Симпсон. — Тут значится, что у женщин, которые принимали лекарство *D*, тоже повысился риск сердечного приступа. Кажется, у меня едет крыша! Получается, лекарство вредно для женщин и вредно для мужчин, но полезно людям в целом».

Вы тоже пришли в недоумение? Если так, вы в хорошей компании. Этот парадокс, который впервые описал реальный статистик по имени Эдвард Симпсон в 1951 году, не давал покоя ученым более 60 лет и продолжает тревожить их до сих пор. Даже в 2016 году, когда я писал эту книгу, вышло четыре новые работы (включая диссертацию), в которых парадокс Симпсона пытались объяснить с четырех разных точек зрения.

В 1983 году Мелвин Новик написал: «Напрашивается вот такой ответ: если мы знаем, что пациент — мужчина или что пациент — женщина, нам не стоит использовать этот препарат. Но если пол неизвестен, препарат лучше использовать! Очевидно, что подобный вывод смехотворен». Я полностью согласен. Считать, что лекарство вредно для мужчин и вредно для женщин, но при этом полезно для людей, было бы действительно смехотворно. Итак, одно из этих трех утверждений должно быть неверным. Но какое? И почему? И как вообще возможна эта путаница?

Чтобы ответить на эти вопросы, нам, конечно, нужно взглянуть на (вымышленные) данные, которые так сильно озадачили нашего доктора Симпсона. Исследование было наблюдательным, а не рандомизированным, с участием 60 мужчин и 60 женщин. Это означает, что пациенты сами решали, принимать препарат или нет. В табл. 8 показано, сколько представителей каждого пола получали препарат *D* и у скольких впоследствии был диагностирован сердечный приступ.

Позвольте мне подчеркнуть, в чем именно заключается парадокс. Как вы можете видеть, 5,0% (1 из 20) женщин в контрольной группе пережила сердечный приступ, в то время как в группе женщин, принявших лекарство, этот показатель составил 7,5%, т.е. лекарство связано с риском сердечного приступа у женщин. У мужчин сердечный приступ случился у 30% в контрольной группе и у 40% в группе принявших

лекарство. Значит, лекарство связано с риском сердечного приступа у мужчин. Доктор Симпсон был прав.

Таблица 8. Вымышленные данные, иллюстрирующие парадокс Симпсона

Пол	Контрольная группа (не принимала лекарство)		Экспериментальная группа (принимала лекарство)	
	Был сердечный приступ	Не было сердечного приступа	Был сердечный приступ	Не было сердечного приступа
Женщины	1	19	3	37
Мужчины	12	28	8	12
Итого	13	47	11	49

Но теперь посмотрите на третью строку таблицы. В контрольной группе сердечный приступ был у 22%, а в группе принявших лекарства — у 18%. Итак, если судить по итогам эксперимента, препарат *D*, похоже, снижает риск сердечного приступа у населения в целом. Добро пожаловать в загадочный мир парадокса Симпсона!

Почти 20 лет я пытаюсь убедить научное сообщество в том, что парадокс Симпсона ставит нас в тупик из-за неправильного применения законов причинности к статистическим соотношениям. Если использовать причинно-следственные обозначения и диаграммы, то можно четко и однозначно решить, предотвращает ли препарат D сердечные приступы или вызывает их. По сути, парадокс Симпсона — это загадка, связанная с конфаундерами, и ее реально решить теми же методами, которые мы уже использовали в похожем случае. Любопытно, что авторы трех из четырех работ 2016 года, о которых я упомянул, продолжают сопротивляться этому решению.

Любая попытка разрешить парадокс (особенно если ему уже несколько десятилетий) должна соответствовать базовым критериям. Во-первых, как я сказал выше в связи с парадоксом Монти Холла, ей следует объяснить, почему люди находят парадокс удивительным или невероятным. Во-вторых, ей нужно показать тип сценариев, в которых возможно его появление. В-третьих, когда парадокс все-таки возникает, и нам надо сделать выбор между двумя правдоподобными, но противоречивыми утверждениями, важно указать, какое из утверждений является правильным.

Давайте начнем с вопроса, почему парадокс Симпсона вызывает удивление. Чтобы ответить на него, надо провести различие между двумя вещами — инверсией Симпсона и парадоксом Симпсона.

Инверсия Симпсона — это чисто числовое явление: как видно из табл. 7, это изменение относительной частоты какого-то события в двух или более различных выборках при объединении выборок. В нашем примере мы увидели, что $3/40 > 1/20$ (частота сердечных приступов среди женщин, принимавших и не принимавших лекарство D) и $8/20 > 12/40$ (частота среди мужчин). Тем не менее, когда мы объединили показатели женщин и мужчин, неравенство изменило направление на противоположное: $(3 + 8) / (40 + 20) < (1 + 12) / (20 + 40)$. Если вы считали такой поворот математически невозможным, то, скорее всего, неверно применяли или неверно запомнили свойства дробей. Многие люди, кажется, считают, что если $A/B > a/b$ и $C/D > c/d$, то $(A + C) / (B + D) > (a + c) / (b + d)$. Но это общее представление ошибочно. Только что приведенный нами пример его опровергает.

Инверсию Симпсона можно обнаружить в наборах данных из реальной жизни. Вот прекрасный образец для фанатов бейсбола, касающийся двух звездных бейсболистов — Дэвида Джастиса и Дерека Джитера. В 1995 году у Джастиса был более высокий средний показатель: 0,253 против 0,250. В 1996 году у Джастиса снова был более высокий средний показатель 0,321 против 0,314. А в 1997 году он набрал больше очков, чем Баттер, третий сезон подряд: 0,329 против 0,291. Тем не менее

за три сезона вместе взятых больше очков оказалось у Джитера! Табл. 8 демонстрирует расчеты для читателей, которые хотели бы их проверить.

Как один игрок может быть хуже, чем другой, в 1995, 1996 и 1997 годах, но лучше в течение трехлетнего периода? Эта инверсия напоминает о лекарстве из нашего примера. На самом деле это невозможно; все дело в том, что мы использовали слишком простое слово («лучше») для описания сложного процесса усреднения по разным сезонам. Обратите внимание, что выходы на биту (знаменатели) не распределяются равномерно по годам. В 1995 году у Джитера было их очень мало, поэтому его довольно низкий средний показатель в этом году мало повлиял на общий средний показатель. Однако у Джастиса было намного больше выходов на биту в его наименее продуктивном году, 1995-м, и это привело к снижению общего среднего показателя. Как только вы поймете, что «лучший нападающий» определяется соперничеством лицом к лицу, а средневзвешенным значением, которое учитывает, как часто играл каждый из них, думаю, все это будет уже не так удивительно.

Таблица 8. Данные (невымышленные), иллюстрирующие инверсию Симпсона

Бейсболист	**Хиты / Выходы на биту**			
	1995	1996	1997	За три года
Дэвид Джастис	104/411 = = 0,253	45/140 = = 0,321	163/495 = 0,329	312/1 046 = = 0,298
Дерек Джитер	12/48 = = 0,250	183/582 = = 0,314	190/654 = = 0,291	385/1 284 = = 0,300

Инверсия Симпсона, конечно же, удивляет некоторых людей и даже фанатов бейсбола. Каждый год у меня появляются студенты, которые сначала не могут поверить в такие вещи. Но потом они идут домой, работают над подобными примерами и утрачивают сомнения. Просто они начинают по-новому, немного глубже понимать, как работают числа

(и особенно агрегированные показатели). Я не называю инверсию Симпсона парадоксом, потому что это по большому счету вопрос исправления ошибочных представлений о том, как ведут себя средние значения. Парадокс — нечто большее: он должен повлечь за собой конфликт между двумя глубоко укоренившимися убеждениями.

У профессиональных статистиков, которые работают с числами каждый день своей жизни, еще меньше оснований считать инверсию Симпсона парадоксом. Простое арифметическое неравенство не могло бы озадачить и увлечь их до такой степени, чтобы они продолжали писать о нем статьи 60 лет спустя.

Вернемся теперь к нашему основному примеру — парадоксу с лекарством. Я объяснил, почему три утверждения («вредно для мужчин», «вредно для женщин» и «полезно для людей), интерпретируемые как увеличение и уменьшение пропорций, не противоречат друг другу математически. И все же вам может показаться, что это физически невозможно. Странно, что лекарство способно вызвать одновременно у меня и у вас сердечный приступ, но в то же время предотвратить сердечный приступ у нас обоих. Это интуитивное чувство универсально; оно появляется у нас в двухлетнем возрасте, задолго до того, как мы начинаем изучать числа и дроби. Поэтому я думаю, вы испытаете облегчение, узнав, что не нужно отказываться от интуиции. Лекарства с такими свойствами пока не изобрели и не изобретут никогда, что мы можем доказать математически.

Первым внимание к этому интуитивно очевидному принципу привлек статистик Леонард Сэвидж. В работе 1954 года он назвал его «верное дело». Он писал: «Бизнесмен задумывается о покупке определенного объекта недвижимости. При этом он учитывает исход следующих президентских выборов. Чтобы прояснить этот вопрос, он спрашивает себя, купил бы он этот объект, если бы знал, что выиграет кандидат-демократ, и приходит к выводу, что да. Потом он задает тот же вопрос о кандидате-республиканце и приходит к такому же выводу. Осознав, что покупка состоялась бы в любом случае, он решается на нее, несмотря на то, что не знает, кто победит. Очень редко решение может быть принято на основе этого принципа,

но… Я не знаю другого экстралогического принципа, управляющего решениями, который было бы так легко принять».

Замечание Сэвиджа в конце цитаты особенно проницательно: он понимает, что принцип верного дела экстралогический. Более того, если интерпретировать его верно, окажется, что он основан на причинно-следственной, а не классической логике. Кроме того, он говорит, что «не знает иного… принципа, который». Очевидно, что он говорил о нем со многими людьми, и они нашли подобное рассуждение очень убедительным.

Чтобы связать принцип верного дела у Сэвиджа с обсуждением выше, предположим, что на самом деле выбор стоит между двумя объектами недвижимости — *A* и *B*. Если победит демократ, у бизнесмена есть 5%-ный шанс заработать доллар на объекте *A* и 8%-ный шанс заработать доллар на объекте *B*. Таким образом, *B* предпочтительнее *A*. Если выиграет республиканец, у него есть 30%-ный шанс заработать доллар на объекте *A* и 40%-ный шанс заработать доллар на объекте *B*. И снова B оказывается предпочтительнее *A*. Согласно принципу верного дела, ему точно нужно купить объект *B*. Но наблюдательные читатели заметят, что числовые величины здесь такие же, как и в истории Симпсона, а значит, покупка объекта *B* может оказаться поспешным решением.

Более того, аргумент, приведенный выше, содержит очевидный недостаток. Если решение бизнесмена купить недвижимость способно повлиять на исход выборов (например, если за его действиями следили СМИ), то покупка недвижимости A окажется в его интересах. А вред от избрания не того президента перевесит любую финансовую выгоду от сделки, когда президент уже будет выбран.

Чтобы принцип верного дела проявил себя, мы должны утвердиться в том, что решение бизнесмена не повлияет на исход выборов. Если бизнесмен уверен, что его решение не окажет воздействия на вероятность победы демократов или республиканцев, он может спокойно покупать недвижимость *B*.

Обратите внимание, что отсутствующий ингредиент (который Сэвидж не указал явно) — предположение о причине. Правильная версия его принципа будет выглядеть так: действие,

которое, по нашему предположению, повышает вероятность некоего результата и в том случае, если событие C произошло, и в том случае, если оно не произошло, повысит его вероятность также и в случае, когда мы не знаем, произошло ли *C*… при условии, что само действие не изменит вероятность *C*. В частности, не существует такого понятия, как «хорошее / хорошее / плохое» лекарство. Этот исправленный вариант принципа Сэвиджа не вытекает из классической логики: чтобы доказать его, понадобится причинное исчисление с привлечением оператора *do*. Наша сильная интуитивная убежденность в невероятности такого лекарства предполагает, что люди (а также машины, запрограммированные на подражание человеческим мыслям) используют что-то вроде *do*-исчисления для направления интуиции.

В соответствии с исправленным принципом, одно из следующих трех утверждений должно быть ложным: препарат *D* повышает вероятность сердечного приступа у мужчин и женщин; препарат *D* снижает вероятность сердечного приступа у населения в целом; препарат не меняет количество мужчин и женщин. Поскольку крайне маловероятно, что лекарство может изменить пол пациента, одно из первых двух утверждений должно быть ложным.

Какое же? Не стоит искать подсказок в табл. 7. Чтобы ответить на этот вопрос, нужно рассмотреть не только данные, но и как они были получены. Как всегда, обсудить этот процесс без диаграммы причинности просто невозможно.

Диаграмма на рис. 44 учитывает важную информацию: препарат не влияет на пол пациента; пол влияет на риск сердечного приступа (у мужчин риск выше); пациент решил принять лекарство *D* или отказался от него. В этом эксперименте женщины явно предпочитали принимать *D*, а мужчины чаще от него отказывались. Таким образом, пол — осложняющая переменная, влияющая и на лекарство, и на сердечный приступ. Чтобы объективно оценить, как лекарство влияет на сердечный приступ, нужно сделать поправку на конфаундер. Для этого надо рассмотреть данные по мужчинам и женщинам отдельно, а затем — взяв среднее значение:

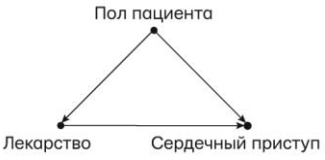


Рис. 42. Диаграмма причинности для примера с парадоксом Симпсона

В группе женщин сердечный приступ случился у 5,0% не принимавших лекарство *D* и у 7,5% принимавших лекарство.

В группе мужчин сердечный приступ случился у 30% принимавших лекарство *D* и у 40% принимавших лекарство.

В среднем (поскольку мужчины и женщины встречаются одинаково часто) сердечный приступ случился у 17,50% не принимавших лекарство *D* (среднее между 5 и 30) и у 23,75% принимавших лекарство (среднее между 7,5 и 40).

Это четкий и недвусмысленный ответ, который мы искали. Лекарство *D* нельзя назвать «плохим / плохим / хорошим» — оно «плохое / плохое / плохое» — для мужчин, женщин и людей в целом.

Я не хочу, чтобы из этого примера у вас создалось впечатление, что агрегировать данные всегда неверно, а разделять их всегда верно. Все зависит от процесса, который произвел данные. В парадоксе Монти Холла мы увидели, что, изменив правила игры, мы также изменили ее исход. Тот же принцип работает и здесь. Я использую другую историю, чтобы продемонстрировать, когда объединение данных окажется уместным. Хотя данные будут абсолютно такими же, роль скрытой третьей переменной изменится, и то же произойдет с результатом.

Давайте начнем с предположения о том, что лекарство B снижает артериальное давление (АД), повышение которого, как известно, может привести к сердечному приступу. Естественно, исследователи лекарства *B* хотели увидеть, не понизит ли оно риск сердечного приступа, поэтому они измеряли артериальное давление пациентов после приема лекарства, а не только фиксировали, был ли у них сердечный приступ.

Табл. 9 показывает данные из исследования лекарства *B*. Она вам знакома: в ней те же показатели, что и в табл. 7. Тем не менее вывод будет абсолютно противоположным.

Таблица 9. Вымышленные данные для примера с артериальным давлением

Артериальное давление	Контрольная группа (не принимала лекарство)		Экспериментальная группа (принимала лекарство)	
	Был сердечный приступ	Не было сердечного приступа	Был сердечный приступ	Не было сердечного приступа
Низкое	1	19	3	37
Высокое	12	28	8	12
Итого	13	47	11	49

Как видите, прием лекарства *B* оказал эффект: в экспериментальной группе давление понизилось у вдвое большего числа человек (у 40 из 60 по сравнению с 20 из 60 в контрольной группе). Другими словами, оно сделало в точности то, что должно делать лекарство против сердечного приступа. Оно вывело людей из категории высокого риска в категорию более низкого риска. Этот фактор перевешивает все остальное, и мы можем прийти к обоснованному выводу о том, что часть табл. 9 с агрегированной информацией дает нам верный результат.

Как обычно, диаграмма причинности все прояснит и позволит нам вывести результат механически, даже не думая о данных и о том, понижает или повышает это лекарство наше кровяное давление. В этом случае скрытая третья переменная — артериальное давление, и диаграмма выглядит, как показано на рис. 43. Здесь артериальное давление — скорее посредник, чем вмешивающийся фактор. Один взгляд на диаграмму показывает, что на взаимосвязь лекарства и сердечного приступа

не действует конфаундер (т.е. нет черного хода), поэтому стратификация данных не требуется. Более того, если учитывать только артериальное давление, мы исключим один из каузальных путей (возможно, основной) для действия лекарства. По обеим этим причинам наш вывод прямо противоположен выводу для препарата *D*: препарат *B* работает и совокупные данные подтверждают этот факт.

С исторической точки зрения примечательно, что Симпсон в статье 1951 года, которая вызвала всю эту шумиху, сделал то же самое, что и я. Он представил две истории с абсолютно одинаковыми данными. В одном примере было интуитивно понятно, что агрегирование данных оказалось, как он выразился, «разумной интерпретацией»; в другом более разумным стало разделение данных. Итак, Симпсон понял, что это парадокс, а не просто инверсия. Однако он не предложил никакого решения, кроме как использовать здравый смысл. И самое важное: он не предположил, что, если история содержит дополнительную информацию, которая позволяет различить «разумное» и «неразумное», возможно, статистикам стоит учесть ее при анализе.

Рис. 43. Диаграмма причинности для примера с парадоксом Симпсона (второй вариант)

Деннис Линдли и Мелвин Новик рассмотрели это предположение в 1981 году, но не смогли примириться с гипотезой, что правильное решение зависит от причинной истории, а не от данных. Они признали: «Мы могли бы использовать язык причинно-следственных связей… Мы решили этого не делать и вообще не обсуждать причинность, потому что, хотя это понятие широко используется, у него как будто нет четкого определения». Так они обобщили фрустрацию пяти

поколений статистиков, которые понимали, что информация о причинно-следственных связях чрезвычайно необходима, но язык для ее выражения безнадежно отсутствует. В 2009 году, за четыре года до смерти в возрасте 90 лет, Линдли признался мне, что он не написал бы приведенные выше слова, если бы моя книга была доступна в 1981 году.

Некоторые читатели моих книг и статей предположили, что правило, регулирующее агрегирование и разделение данных, основывается на временном приоритете в обработке и скрытой третьей переменной. Они утверждают, что в случае с артериальным давлением данные необходимо агрегировать, потому что измерение давления происходит после того, как пациент принимает лекарство, но в случае с полом данные нужно стратифицировать, потому что пол пациента определен заранее. Хотя это правило работает во многих случаях, его нельзя считать универсальным. Простой случай — *M*-тип (игра 4 в главе 4). Здесь *B* может предшествовать *A*; тем не менее мы все равно не должны ставить условие на *B*, потому что это нарушит критерий черного хода. Мы должны обратиться к причинно-следственной структуре рассказа, а не к временной информации.

Наконец, вы можете задаться вопросом, возможен ли парадокс Симпсона в реальном мире. Ответ будет положительным. Конечно, он встречается недостаточно часто, чтобы статистики наблюдали его ежедневно, однако он не совсем неизвестен и, вероятно, происходит чаще, чем об этом сообщают статьи в научных журналах. Вот два задокументированных случая.

Наблюдательное исследование, опубликованное в 1996 году, показало, что открытая операция по удалению камней в почках чаще завершалась успехом, чем эндоскопическая операция, которую, делали при небольших камнях. Кроме того, оно продемонстрировало, что, если камни в почках были больше, повышался и процент успеха. Но в целом для открытой операции он был ниже. Как и в нашем первом примере, выбор метода лечения зависел от состояния пациента: если камни были крупнее, открытая хирургическая операция была вероятнее, но прогноз оказывался хуже.

В исследовании заболеваний щитовидной железы, опубликованном в 1995 году, курильщики продемонстрировали более высокий коэффициент выживания (76%) в течение 20 лет, чем некурящие участники (69%). Но у некурящих этот показатель был выше в шести из семи возрастных групп, а в седьмой разница оказалось минимальной. Фактор возраста явно повлиял и на курение, и на выживание: средний курильщик был моложе среднего некурящего (возможно, потому что курильщики старшего возраста уже умерли). Разделив данные по возрастным группам, мы пришли к выводу, что курение отрицательно влияет на выживание.

Поскольку парадокс Симпсона настолько плохо поняли, некоторые статистики специально стараются его избежать. Слишком часто они пытаются бороться с симптомом, инверсией Симпсона, ничего не делая с болезнью — конфаундерами. Вместо того чтобы подавлять симптомы, необходимо обращать на них внимание. Парадокс Симпсона предупреждает нас о случаях, когда по крайней мере один из статистических трендов (в агрегированных данных, разделенных или в тех и других) не может отражать причинно-следственное влияние. Есть, конечно, и другие тревожные знаки. Если оценить причинно-следственный эффект в совокупности, получившаяся величина, к примеру, может оказаться выше, чем каждая аналогичная величина в каждой страте. И снова приходится повторять: этого не должно произойти, если мы должным образом учли вмешивающиеся факторы. Однако по сравнению с такими признаками инверсию Симпсона труднее игнорировать именно потому, что это разворот, качественное изменение знака. Идея вредного / вредного / полезного препарата вызовет недоверие даже у трехлетнего ребенка — и совершенно справедливо.

Парадокс Симпсона в картинках

До сего момента большая наши примеры инверсии и парадокса Симпсона в основном включали двоичные переменные: пациент либо принимал Лекарство *D*, либо нет, и либо пере-

живал сердечный приступ, либо нет. Однако инверсия может возникнуть и с непрерывными переменными, и, возможно, в этом случае она будет понятнее за счет возможности ее проиллюстрировать.

Представьте исследование, в котором измеряют еженедельную физическую активность и уровень холестерина у людей разных возрастных групп. Если нанести количество часов, потраченных на физкультуру, на ось *x* и уровень холестерина на ось *Y*, как показано на рис. 44 (a), то для каждой возрастной группе наблюдается наклон вниз, и это, вероятно, означает, что физическая активность понижает уровень холестерина. С другой стороны, если использовать ту же диаграмму рассеяния, но не делить данные по возрастам, как на рисунке 44 (*b*), то мы увидим выраженную направленность вверх, которая говорит: чем больше люди занимаются физкультурой, тем выше их уровень холестерина. Ситуация с плохим-плохим-хорошим лекарством повторяется, только в его роли выступает Физическая активность. Кажется, что она положительно влияет на каждую возрастную группу, но вредит населению в целом.

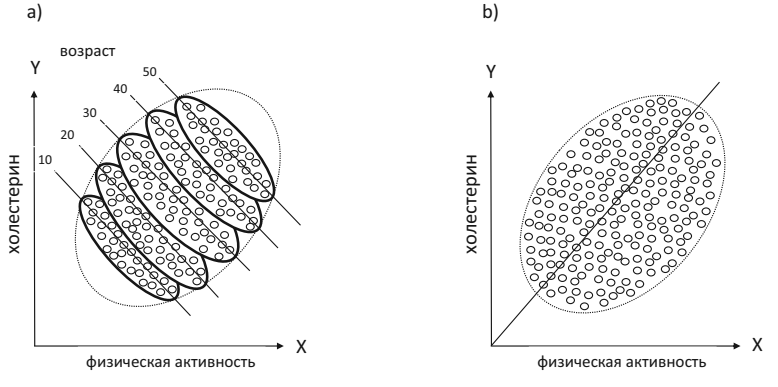


Рис 44. Парадокс Симпсона: физическая активность представляется полезной (направление вниз) в каждой возрастной группе, но вредной (направление вверх) в популяции в целом.

Чтобы решить, полезна или вредна физическая активность, мы, как всегда должны обратиться к истории, которая стоит

за данными. Данные показывают, что люди старшего возраста в нашей популяции больше занимаются физическими упражнениями. Поскольку ситуация, когда Возраст служит причиной Физической активности представляется более вероятной, чем обратная, и поскольку Возраст способен оказывать каузальное воздействие на Холестерин, мы приходим к выводу, что Возраст может быть осложнителем для Физической активности и Холестерина. Значит, нужно сделать корректировку по Возрасту. Другими словами, следует посмотреть на данные, распределенные по возрастам, и мы увидим, что физические упражнения приносят пользу, независимо от возраста.

Родственник парадокса Симпсона десятилетиями мелькал в литературе по статистике, и его легко интерпретировать визуальными средствами. Фредерик Лорд изначально сформулировал этот парадокс в 1967 году. И снова он вымышленный, но вымышленные примеры (вроде мысленных парадоксов Эйнштейна) всегда обеспечивают хороший способ нащупать границы нашего понимания.

Лорд наблюдает за университетом, администрация которого хочет понять, как питание, предлагаемое студентам в столовых, влияет на их вес — в частности, оказывает ли оно разный эффект на юношей и девушек. С этой целью студенты сначала взвешиваются в сентябре, а потом в следующем июне. На рис. 45 результаты представлены на графике, где эллипсы снова отражают диаграммы рассеяния данных. Потом университет приглашает двух статистиков, они рассматривают полученные данные и приходят к противоположным выводам.

Первый статистик смотрит на распределение веса у девушек в целом, и отмечает, что их средний вес в июне остался таким же, как в сентябре. (Это показывает симметрия диаграммы рассеяния вдоль линии $W_F = W_I$, то есть, конечный вес = исходный вес). Естественно, отдельные девушки могут набирать или терять вес, но в целом прибавка равна нулю. То же справедливо и для юношей. Из этого статистик делает вывод, что рацион не влияет на два пола по-разному.

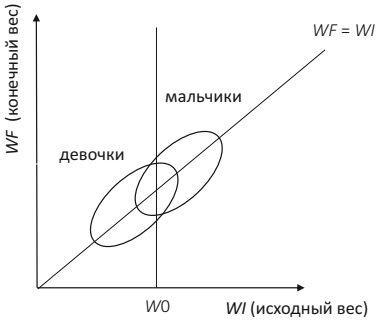


Рис 45. Парадокс Лорда. (Эллипсы представляют диаграммы рассеяния). В целом, ни девушки, ни юноши не прибавляют в весе в течение года, но в каждой страте с одинаковым исходным весом юноши имеют тенденцию поправляться больше девушек.

Второй статистик, со своей стороны, утверждает, что, поскольку на конечный вес студентов сильно влияет исходный вес, необходимо стратифицировать их по исходному весу. Если сделать вертикальный срез, пройдя через оба эллипса, что позволит посмотреть только на юношей и девушек с определенным исходным весом (например, W_0 на рис. 45), вы заметите, что вертикальная линия пересекает эллипс «Юноши», выше, чем эллипс «Девушки», хотя есть некоторое наложение. Это значит, что у юношей, которые начали с веса W_0, конечный вес (W_F) в среднем будет выше, чем у девушек которые начали с веса W_0. Соответственно, пишет Лорд, «второй статистик приходит к выводу, как это обычно бывает в таких случаях, что юноши демонстрируют существенно большую прибавку, чем девушки, если соответствующим образом учесть разницу в исходном весе между полами».

Что же делать университетскому диетологу? Лорд пишет: «Выводы обоих статистиков с виду верны». То есть, не нужно заниматься подсчетами, чтобы увидеть: два веских аргумента ведут к двум разным выводам. Достаточно посмотреть на рисунок: на рис. 45 мы видим, что юноши в каждой страте (в любом вертикальном сечении) поправляются больше, чем девушки. В то же время, очевидно, что ни юноши, ни девушки в итоге

не прибавили ничего. Как это возможно? Разве общая прибавка веса — не средний показатель для прибавки в каждой группе?

Теперь, когда мы прекрасно разбираемся в тонкостях парадокса Симпсона и принципа «верного дела», мы знаем, что не так с этим аргументом. Принцип «верного дела» работает только в случаях, когда относительная доля любой подгруппы (группы с одинаковым весом) остается неизменной. Да, в случае Лорда «воздействие» (пол) сильно влияет на процент студентов в каждой группе с одним весом.

То есть, мы не можем полагаться на принцип «верного дела», и это возвращает нас к началу. Кто же прав? Существует ли разница между средней прибавкой в весе между девушками и юношами, если соответствующим образом учесть разницу в исходном весе между полами? Вывод Лорда весьма пессимистичен: «Обычные исследование такого типа пытается ответить на вопрос, на который просто невозможно дать точный ответ на основе имеющихся данных». Пессимизм Лорда вышел за пределы статистики и привел к изобилию довольно пессимистичных работ по эпидемиологии и биостатистике о том, как сравнивать группы с разной исходной статистикой.

Сейчас я покажу вам, почему пессимизм Лорда не обоснован. На вопрос диетолога можно ответить точно, и, как обычно, надо начать с диаграммы причинности, как на рис. 46. На этой диаграмме мы видим, что Пол (S) становится причиной исходного веса (W_I) и конечного веса (W_F). Кроме того, W_I влияет на W_F независимо от гендера, потому что студент любого гендера, который весит больше в начале года, чаще весит больше и в конце года, что показано на диаграммах рассеяния на рис. 45. Все эти допущения о причинности основаны на здравом смысле; я не думаю, что Лорд не согласился бы с ними.

Интересующая Лорда переменная — прибавка в весе, которая показана на этой диаграмме как Y. Заметьте, что Y относится к W_I и W_F чисто математически, детерминированным образом: $Y = W_F - W_I$. Это значит, что корреляции между Y и W_I (или Y и W_F) равна –1 (или 1), и я показал эту информауию на диаграмме с коэффициентами –1 и +1.

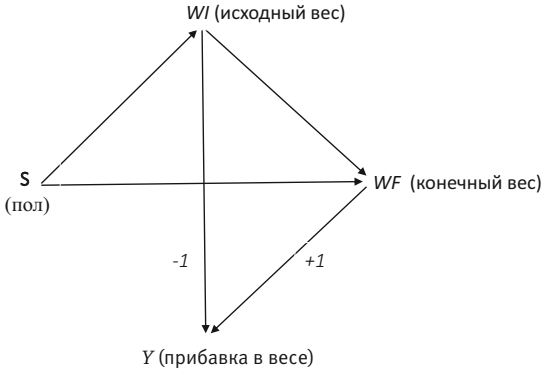


Рис 46. Диаграмма причинности для парадокса Лорда.

Первый статистик просто сравнивает разницу в прибавке веса между юношами и девушками. Между *S* и *Y* не нужно блокировать «черные ходы», а значит, наблюдаемые обобщенные данные дают ответ: эффекта нет, и таков вывод первого статистика.

В то же время, вопрос, на который пытается ответить второй статистик (то есть, «верно сформулированный запрос», описанный во Вступлении) трудно даже сформулировать. Он хочет убедиться, что «разница в исходном весе между полами соответствующим образом учтена» — такие формулировки обычно используют, если хотят сделать поправку по осложнителю. Но W_I не выступает осложнителем для *S* и *Y*. На самом деле это переменная-медиатор, если считать Пол «воздействием». Таким образом, запрос, на который отвечают, вводя ограничение по W_I, нельзя интерпретировать с точки зрения обычной причинности. Такое ограничение в лучшем случае позволит увидеть «прямой эффект» гендера на вес, что мы обсудим в Главе 9. Тем не менее, представляется маловероятным, что второй статистик имел это в виду; скорее всего, он действовал по привычке. И в то же время, этот аргумент позволяет очень легко попасть в ловушку: «Разве общая прибавка — не среднее между прибавками во всех группах»? Нет, если сами группы меняются в зависимости от воздействия! Не забывайте, что

воздействием здесь служит Пол, а не Рацион, и Пол определенно меняет соотношение студентов в каждой страте W_I.

Этот последний комментарий затрагивает еще одну любопытную деталь о парадоксе Лорда в его изначальной формулировке. Хотя изначальным намерением университетского диетолога было «определить эффект рациона» Лорд в своей изначальной статье нигде не говорит о контрольном рационе. Соответственно, мы ничего не можем сказать об эффекте от рациона. Говард Вайнер и Лиза Браун пытаются исправить этот недостаток в статье, вышедшей в 2006 году. Они меняют историю таким образом, что интересующем показателем является влияние рациона (а не пола) на прибавку в весе, в то время как разница между юношами и девушками не учитывается. В их варианте студенты едят в одной из двух столовых с разными рационами. Соответственно, два эллипса на рис. 45 представляют две столовые, в которых предлагают разное питание, что обозначено на рисунке 47 (a). Обратите внимание на то, что студенты, которые в начале весят больше, чаще предпочитают столовую *B*, в то время как те, кто весит меньше, едят в столовой *A*.

Парадокс Лорда теперь вырисовывается с большей ясностью, поскольку запрос точно определен как воздействие рациона на прибавку веса. Первый статистик утверждает, исходя из соображений симметрии, что переход с Рациона A на Рацион B не оказал бы эффекта на прибавку веса (разница $W_F - W_I$ одинаково распределена в обоих эллипсах). Второй статистик сравнивает итоговые показатели веса при Рационе A с показателями для рациона B для группы студентов, которые начали с веса W_0 и приходит к выводу о том, что студенты, получающие рацион *B*, сильнее прибавляют в весе.

Как и раньше, данные (рис. 47 [a]) не скажут вам, кому верить, — именно к такому выводу приходят Вайнер и Браун. Однако диаграмма причинности (рис. 47 [*b*]) поможет разрешить этот вопрос. Между рис. 46 и рис. 47 (*b*) есть существенные различия. Во-первых, первичной величиной становится Рацион, а не Пол. Во-вторых, стрелка, которая изначально указывала от *S* к W_I, теперь меняет направление: исходный вес теперь влияет на рацион, поэтому стрелка направлена от W_I к *D*.

На этой диаграмме W_I — осложнитель для D и W_F, а не медиатор. Таким образом, второй статистик в этом случае будет однозначно прав. Ограничение по исходному весу необходимо, чтобы D и W_F (а также D и Y) освободились от осложнителя. Первый статистик оказался бы неправ, потому что измерял бы только статистические ассоциации, а не причинно-следственные эффекты.

Подводя итог, скажем, что для нас главный урок парадокса Лорда в том, что он не более парадоксален, чем парадокс Симпсона. В одном из них ассоциация становится обратной, а в другой исчезает. И в обоих случаях диаграмма причинности подскажет, какую процедуру нужно использовать. Однако статистикам, обученным «традиционной» (т.е. не учитывающей модели) методологии и избегающим оптики причинности, представляется глубоко парадоксальным тот факт, что вывод, верный в одном случае, будет неверным в другом, при том, что данные выглядят совершенно одинаково.

Теперь, хорошо проработав коллайдеры, осложнители и опасности, которыми они грозят, мы наконец-то готовы пожать плоды нашего труда. В следующей главе мы начнем подъем по Лестнице причинности, начав со второго уровня — интервенции.

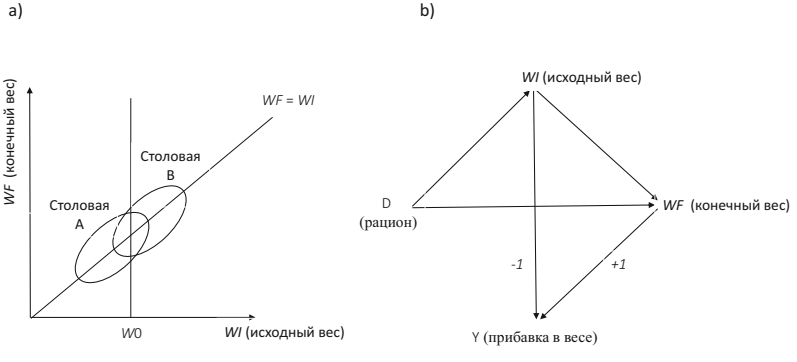


Рис 47. Обновленная версия парадокса Лорда по Вайнеру и Браун и соответствующая диаграмма причинности.

Глава 7
За пределами поправок: покорение горы интервенции

В том, у кого боязнь согрешить
проявляется прежде, чем мудрость,
мудрость укрепится;
утратит же ее тот,
у кого она проявляется прежде,
чем боязнь согрешить Авот. 3:9
Раби Ханина бен Доса (I век н.э.)

В этой главе мы, наконец, храбро предпримем восхождение на второй уровень Лестницы Причинности, уровень интервенций — святой Грааль каузального мышления с древнейших времен до наших дней. Этот уровень задействован в попытках предсказать эффекты еще не испытанных действий и стратегий, от способов лечения до социальных программ, от экономической политики до личного выбора. Конфаундеры были основным препятствием, заставлявшим нас путать *наблюдаемое с осуществляемым*. Удалив это препятствие с помощью техники блокирования путей и критерия черного хода, мы можем картировать путь к горе Интервенции с систематической точностью. Для начинающего альпиниста самые безопасные тропы наверх — это поправки черного хода и различные родственные им техники, некоторые из них упомянуты тут в рубриках «Поправки парадного входа» и «Инструментальные переменные».

Однако не во всех случаях эти пути доступны, поэтому опытному скалолазу эта глава предоставляет универсальный инструмент картирования, так называемое *do*-исчисление, позволяющее исследователю обследовать и нанести на карту все пути на вершину Интервенции, как бы извилисты они ни были. Если путь зафиксирован на карте и все тросы и карабины наготове, наше восхождение на вершину обречено увенчаться успехом!

Самый простой путь: формула поправки черного хода

Для многих исследователей самый (или единственный) известный метод предсказания эффектов интервенции — поправки по конфаундерам по соответствующей формуле. Этот метод разумно использовать, когда вы уверены, что у вас есть данные по достаточному набору переменных (снимающих осложнения), чтобы заблокировать все черные ходы между интервенцией и результатом. Для этого мы должны измерить средний каузальный эффект интервенции, вначале оценив ее эффект на каждом уровне или страте, снимающих осложнение переменной. Затем мы исчисляем среднее взвешенное этих страт, где каждая из них определена в соответствии со своим распространением в популяции. Если, например, переменная, по которой вводится поправка, — это пол, мы прежде всего оцениваем каузальный эффект для мужских и женских особей отдельно, затем усредняем его, если в популяции, как чаще всего бывает, соотношение полов один к одному. Если соотношения иные, скажем особей мужского пола — $\frac{2}{3}$, а женского — $\frac{1}{3}$, тогда для оценки среднего каузального воздействия нужно взять соответствующим образом средние взвешенные.

Роль, которую в этой процедуре играет критерий черного хода, — это гарантия, что каузальный эффект в каждой страте переменной, снимающей осложнения, не что иное, как наблюдаемый в этой страте тренд. Таким образом, каузальный эффект можно вывести из данных по частям, страта за стра-

той. В отсутствие критерия черного хода у исследователей нет гарантии, что поправки оправданы.

Пример с вымышленным лекарством в главе 6 — самая простая из возможных ситуаций: одна экспериментальная переменная (лекарство *D*), один исход (инфаркт), один конфаундер (пол) и все три переменные бинарны. Этот пример демонстрирует, как мы получаем среднее взвешенное по условным вероятностям *P (инфаркт | лекарство)* в каждой из страт (пол). Но описанную выше процедуру легко модифицировать так, чтобы она годилась и для более сложных ситуаций, включая множественность конфаундеров и множественность страт.

Однако во многих случаях переменные *X*, *Y* или *Z* принимают численные значения: доход, или рост, или вес при рождении. Мы наблюдали это в визуальном образце с парадоксом Симпсона. Поскольку переменная способна принимать (по крайней мере, для всех практических целей) бесконечное множество возможных значений, мы не в состоянии перечислить их все в таблице, как было сделано в главе 6.

Очевидное решение — распределить численные значения переменной по конечному и удобному в использовании числу категорий. В таком решении нет ничего принципиально неправильного, однако выбор числа категорий оказывается несколько произвольным. Намного хуже, когда переменных, по которым вводятся поправки, оказывается достаточно много, число категорий растет по экспоненте, что делает исчисление по этой процедуре затруднительным; еще хуже, что во многих стратах при этом нет ни одного образца и они не могут, таким образом, дать оценку вероятности.

Статистики изобрели хитроумные методы избавления от этой проблемы «проклятья множественных измерений». В большинстве из них в том или ином виде применяется экстраполяция, когда для данных подбирается соответствующая им гладкая функция, с помощью которой закрываются дыры, оставленные пустыми стратами.

Наиболее часто из всех гладких функций используется, конечно, линейное аппроксимирование; все XX столетие оно честно служило рабочей лошадкой в большей части работ,

связанных с количественным исчислением, в науках об обществе и поведении. Мы уже видели, как Сьюалл Райт погрузил свои путевые диаграммы в контекст линейных уравнений, и отметили одно преимущество, которое дает это погружение: каждое каузальное воздействие может быть представлено одним числом (путевым коэффициентом). Второе и не менее важное преимущество линейных аппроксимаций — невероятная простота подсчета поправочной формулы. Ранее мы познакомились с изобретенной Фрэнсисом Гальтоном линией регрессии, когда берется облако точек данных и через это облако интерполируется прямая, наиболее соответствующая их распределению. В случае одной экспериментальной (независимой) переменной (X) и одной зависимой (Y) уравнение для линии регрессии выглядит так: $Y = aX + b$. Параметр a (часто обозначаемый как r_{YX}, коэффициент регрессии Y на X) рассказывает нам о наблюдаемой в среднем тенденции: увеличение X на 1 приведет в среднем к увеличению Y на a единиц. Если у X и Y нет конфаундеров, мы можем использовать это выражение как нашу оценку интервенции по увеличению X на 1. Но что же происходит, если имеется конфаундер, Z? В этом случае коэффициент корреляции r_{YX} не сообщает нам средний каузальный эффект: он передает нам только среднюю наблюдаемую тенденцию. В этом была загвоздка у Райта в случае проблемы веса морских свинок при рождении, обсужденной в главе 2: очевидная прибавка в весе (5,66 грамма) за дополнительный день беременности была смещенной оценкой, потому что осложнялась эффектом меньшего размера помета. Но выход все же есть: разместить все данные по трем переменным так, чтобы каждое значение (X, Y, Z) соответствовало одной точке в пространстве в одной системе координат. В этом случае данные образуют облако точек в XYZ-пространстве. Аналогом линии регрессии здесь будет плоскость регрессии, описываемая уравнением $Y = aX + bZ + c$. Мы с легкостью вычислим a, b и c из этих данных. В этот момент происходит нечто замечательное, о чем Гальтон не догадывался, а Карл Пирсон и Джордж Удни Юл знали точно. Коэффициент a теперь дает нам коэффициент регрессии Y на X уже с поправкой по Z (он называется коэффи-

циентом частичной регрессии и записывается как $r_{YX.Z}$). Таким образом, мы можем избежать трудоемкой процедуры подсчета регрессии *Y* на *X* для каждого уровня *Z* и исчисления среднего взвешенного для этих коэффициентов регрессии. Природа сама все усредняет за нас! Нам нужно только рассчитать плоскость, лучше всего описывающую наши данные. Статистические пакеты справляются с этим моментально. Коэффициент a в уравнении этой плоскости, $Y = aX + bZ + c$, автоматически вносит поправку в наблюдаемый тренд *Y* на *X* по конфаундеру *Z*. Если *Z* — единственный конфаундер, то a — это среднее каузальное воздействие *X* на *Y*. Поистине чудесное упрощение!

Эта процедура также легко расширяется для работы со многими переменными. Если набор переменных *Z* удовлетворяет критерию черного хода, тогда коэффициент при *X* в уравнении регрессии a оказывается не чем иным, как средним каузальным воздействием *X* на *Y*.

По этой причине поколения исследователей верили, что коэффициенты регрессии после введения поправок (иначе — коэффициенты частичной регрессии) каким-то образом наделены каузальной информацией, которой нет в коэффициентах регрессии без поправок. Ничего не может быть дальше от истины. Коэффициенты регрессии, с поправками или без, — это только статистические тенденции, и в них самих по себе каузальная информация не содержится. Коэффициент $r_{YX.Z}$ представляет собой каузальное воздействие *X* на *Y*, а r_{YX} — нет исключительно потому, что у нас есть диаграмма, показывающая, что *Z* — это конфаундер для *X* и *Y*.

Короче говоря, иногда коэффициент регрессии представляет собой каузальное воздействие, иногда нет, но для того, чтобы понять разницу, недостаточно одних только данных. Для вооружения $r_{YX.Z}$ причинностной легитимностью нужны еще два ингредиента. Во-первых, путевая диаграмма должна представлять собой правдоподобную картину реальности, и во-вторых, переменные, по которым вводятся поправки, должны соответствовать критерию черного хода.

Вот поэтому проводимое Сьюаллом Райтом разграничение между путевыми коэффициентами (представляющими собой

каузальные воздействия) и коэффициентами регрессии (представляющими собой тенденции в распределении данных) было таким принципиальным. Путевые коэффициенты отличаются от коэффициентов регрессии фундаментальным образом, хотя первые часто выводятся из последних. Ни Райту, однако, ни всем, кто занимался эконометрией и путевым анализом после него, не довелось узнать, что его вычисления были неоправданно сложны. Он мог бы получить путевые коэффициенты из коэффициентов частичной корреляции, если бы только знал, что правильный набор переменных, по которым нужна поправка, легко вывести из самой путевой диаграммы.

Следует помнить также, что поправки, основанные на регрессии, работают только для линейных моделей, что означает значительные допущения при выборе модели. В случае линейных моделей мы теряем возможность передавать нелинейные взаимодействия, например, когда воздействие X на Y зависит от уровня Z. В свою очередь, поправки черного хода нормально работают даже тогда, когда мы не представляем, какие функции стоят за стрелочками на диаграмме. Однако в этом так называемом непараметрическом случае, нам придется применять другие методы экстраполяции, для того чтобы избавиться от проклятья многомерности.

Подводя итоги, отметим, что формула поправок черного хода и критерий черного хода как две стороны одной монеты. Критерий черного хода сообщает нам, какие переменные следует использовать, чтобы снять осложнения. Формула поправок непосредственно снимает их. В простейшем случае линейной регрессии коэффициенты частичной регрессии осуществляют поправку черного хода имплицитно. В непараметрических случаях нам придется выполнять поправки эксплицитно либо с помощью формулы поправок черного хода прямо с исходными данными, либо с какой-либо их экстраполированной версией.

Вы могли подумать, что наше восхождение на гору Интервенцию на этом закончилось полнейшим успехом. Однако, к сожалению, поправки не работают совсем, если имеется путь через черный ход, который мы не в состоянии заблокировать, потому что у нас нет требующихся для этого данных. Однако даже в этом случае мы можем использовать определенные при-

емы. Далее я расскажу вам об одном из моих любимых методов, называемом поправкой парадного входа. Хотя он был описан более 20 лет назад, только горстка исследователей за это время воспользовалась этой короткой дорогой на гору Интервенцию, и я убежден, что его потенциал еще предстоит раскрыть.

Критерий парадного входа

Дебаты о каузальном воздействии курения происходили по крайней мере за два поколения до того, как каузальные диаграммы могли бы в них поучаствовать. Мы уже рассмотрели, как неравенство Корнфилда помогло уверить исследователей, что ген курильщика, или конституциональная гипотеза, — очень неправдоподобное предположение. Однако более радикальный подход с использованием каузальных диаграмм пролил бы больше света на гипотетический ген и, вероятно, полностью исключил его из дальнейшего обсуждения.

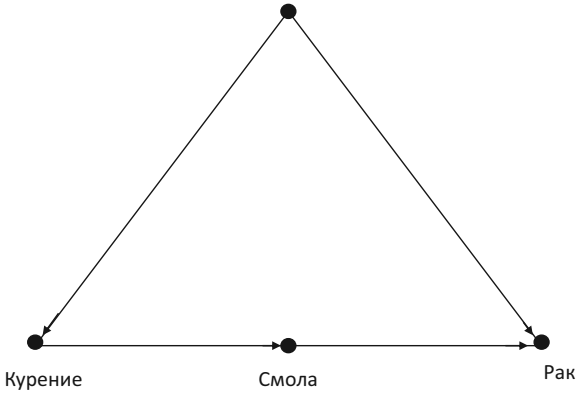


Рис. 41. Гипотетическая каузальная диаграмма для связи курения и рака легких, подходящая для поправок парадного входа

Предположим, что исследователи прошлого смогли измерить отложения смол в легких курильщиков. Еще в 1950-х это называлось в качестве одной из промежуточных стадий в развитии рака легких. Предположим также, что мы, совсем

как министр здравоохранения, хотим исключить гипотезу Р. Э. Фишера о том, что ген курильщика является конфаундером по отношению к привычке курить и раку легких. Тогда ситуацию выше описывает каузальная диаграмма на рис. 41.

Рисунок 41 включает два очень важных допущения, которые будут важны для целей нашего примера. Первое: ген курильщика не оказывает никакого воздействия на формирование отложений смол, которые зависят исключительно от физического действия сигаретного дыма (это допущение отражено на схеме отсутствием стрелки между геном курильщика и смолой; оно, однако, не исключает случайные факторы, не имеющие отношения к гену курильщика). Второе значительное допущение состоит в том, что курение ведет к раку только через накопления отложений смол. Таким образом, на схеме нет прямой стрелки от курения к раку и нет также других непрямых путей.

Допустим, что мы выполняем исследование на основе наблюдаемых данных и собрали информацию по курению, смоле и раку для каждого из участников. Нам, к сожалению, недоступны данные по гену курильщика, потому что неизвестно, существует ли такой ген. Поскольку таким образом у нас отсутствуют сведения по переменной-конфаундеру, мы не в состоянии заблокировать путь через черный ход *курение ← ген курильщика → рак*. Таким образом, мы не можем и использовать поправки черного хода для устранения воздействия конфаундера. Поэтому нам придется искать другие способы. Вместо перемещения черным ходом мы пойдем через парадный вход! В приведенном случае это прямой каузальный путь *курение → смола → рак*, для которого у нас есть данные по всем трем переменным. Интуитивно мы рассуждаем следующим образом. Прежде всего, мы в состоянии оценить средний каузальный эффект влияния курения на смолу, потому что на схеме нет незаблокированных путей через черный ход от курения к раку — путь *курение ← ген курильщика → рак ← смола* уже заблокирован схождением у переменной *рак*. Поскольку этот путь уже заблокирован, нам даже не нужна поправка черного хода. Мы просто наблюдаем вероятности P (*смола* | *привычка курить*) и P (*смола* | *отсутствие привычки*

курить), а разница между ними и будет средним каузальным воздействием курения на смолу. Аналогично диаграмма позволяет нам оценить среднее каузальное воздействие смолы на рак. Чтобы сделать это, мы заблокируем путь черного хода от смолы к раку: *смола* ← *курение* ← *ген курильщика* → *рак*, введя поправки по курению. Здесь пригодятся уроки главы 4: нам нужны только данные по минимальному достаточному набору переменных, снимающих осложнения (здесь — курение). Тогда формула поправки черного хода даст нам вероятности P (*рак* | *do* (*смола*)) и P (*рак* | *do* (*отсутствие смолы*)). Разница между этими двумя вероятностями и будет средним каузальным воздействием смолы на рак.

Теперь нам известно среднее увеличение вероятности отложения смол благодаря курению и среднее увеличение вероятности заболеть раком из-за отложения смол. Можем ли мы как-либо объединить эти вероятности, чтобы получить средний рост заболеваемости раком из-за курения? Да, можем. Рассуждаем мы при этом таким образом: рак возникает двумя путями: при отложении смол и без отложения смол. Если мы заставим кого-либо курить, вероятности этих двух состояний будут соответственно P (*смола* | *do* (*курение*)) и P (*отсутствие смолы* | *do* (*отсутствие курения*)). Однако, если возникнет состояние отсутствия смолы, вероятность рака будет P (*рак* | *do* (*отсутствие смолы*)). Оценив оба сценария по их относительным вероятностям при *do* (курение), получится рассчитать общую вероятность возникновения рака по причине курения. Те же аргументы действуют, если мы не даем кому-либо курить, — *do* (*отсутствие курения*). Разница между результатами дает нам среднее каузальное воздействие курения по сравнению с воздержанием от него на возникновение рака. Как я только что объяснил, мы оцениваем каждую из двух *do*-вероятностей, обсужденных выше, прямо из данных, т.е. записываем их математически в терминах вероятностей, не использующих оператор *do*. Таким образом, математика делает для нас то, чего не могли добиться десятилетия споров и свидетельств конгрессов, — количественно оценить каузальное

воздействие курения на рак, конечно, при условии, что наши предположения верны.

Процесс, который я только что представил, описывающий вероятность P (*рак* | (*курение*)) в терминах вероятностей, исключающих оператор *do*, называется поправкой парадного входа. От поправки черного хода он отличается тем, что мы вносим поправки для двух переменных (курение и смола) вместо одной, и эти переменные лежат на прямом пути от курения к раку, а не на пути через черный ход. Для читателей, знакомых с математическим языком, я покажу эту формулу, которой нет в обычных учебниках статистики. Здесь X — это курение, Y — рак, Z — смола, а U (которое подозрительно отсутствует в формуле) — это ненаблюдаемая переменная, *ген курильщика*:

$$\mathrm{P}(Y|do(X)) = \sum_Z \mathrm{P}(Z=Z,X) \sum_Z \mathrm{P}(Y|X=X,Z=Z)\,\mathrm{P}(X=X). \quad (2)$$

Читателям со вкусом к математике будет интересно сравнить эту формулу с формулой для поправки черного хода, которая записывается так:

$$\mathrm{P}(Y|do(X)) = \sum_Z \mathrm{P}(Y|X, Z=Z)\,\mathrm{P}(Z=Z). \quad (3)$$

Даже для читателей, совсем не владеющих математическим языком, можно сделать несколько интересных замечаний об уравнении (2). Первое и самое важное: в нем нигде нет переменной U (*ген курильщика*). Весь ее смысл как раз в этом. Мы успешно сняли осложнения по U, не обладая никакими данными по ней. Для любого статистика поколения Фишера это выглядело бы как самое настоящее чудо. Во-вторых, в самом начале, во введении, я рассказывал про эстиманд как способ вычислить интересующую нас величину в рамках данного вопроса. Уравнения (2) и (3) — самые сложные и интересные эстиманды в этой книге. Левая сторона представляет вопрос «Каково воздействие X на Y?» Правая сторона — это эстиманд, способ ответа на заданный вопрос. Обратите внимание, что эстиманд не содержит никаких *do*, только see, представленные

вертикальными чертами, и это означает, что его можно рассчитать по имеющимся данным.

К этому моменту, я уверен, многие читатели гадают, насколько этот вымышленный сценарий близок к реальности. Неужели жаркий спор о курении и раке разрешился благодаря одной работе на основе наблюдений и одной каузальной диаграмме? Если мы предположим, что рис. 41 точно отражает причинностный механизм возникновения рака, ответом будет абсолютное «да». Однако то, насколько наши допущения справедливы для реального мира, требует дополнительного обсуждения.

Дэвид Фридман, мой старый друг, занимающийся статистикой в Калифорнийском университете в Беркли, серьезно раскритиковал меня по этому вопросу. Он утверждает, что модель на рис. 41 нереалистична по трем причинам. Во-первых, если ген курильщика» существует, он должен влиять и на то, как тело избавляется от чужеродных веществ в легких, и, таким образом, люди с этим геном будут более склонны к возникновению отложений смол, а люди, лишенные его, — более устойчивы к нему. Поэтому он бы нарисовал стрелку от гена курильщика к смоле, и в этом случае формула парадного входа окажется непригодной. Фридман считает также маловероятным, чтобы курение влияло на возникновение рака только через отложения смол. С уверенностью можно предположить и другие механизмы: не исключено, что курение ведет к хроническому воспалению, которое, в свою очередь, способствует развитию рака. Наконец, говорит он, отложения смол в легких живого человека все равно нельзя измерить со сколь-либо приемлемой точностью, поэтому предложенную мной работу на основе наблюдений не провести в реальном мире.

Я не возражаю против критики Фридмана в этом конкретном примере. Я не специалист по раку, и мне всегда придется оставлять на усмотрение эксперта в данном вопросе, насколько подобная диаграмма адекватно отражает процессы, происходящие в реальном мире. На самом деле одно из самых значительных достижений метода каузальных диаграмм в том,

что они делают допущения прозрачными и открытыми для обсуждения экспертами и политиками.

Тем не менее цель моего примера была не в том, чтобы предложить новый механизм для воздействия курения на организм, а в том, чтобы продемонстрировать, как математика в определенной ситуации способна устранить воздействие конфаундеров, даже если данных по самому конфаундеру нет. Подобную ситуацию легко распознать. В ситуациях, когда каузальное воздействие *X* на *Y* осложняется одним набором переменных (*C*) и опосредуется другим (*M*) (рис. 42) и, более того, опосредующие переменные защищены, как щитом, от воздействий *C*, вы всегда можете оценить воздействие *X*, пользуясь наблюдаемыми данными. Узнав об этом факте, ученым было бы разумно, столкнувшись с неустранимыми конфаундерами, искать защищенные медиаторы. Как говорил Луи Пастер, «удача сопутствует подготовленному уму».

К счастью, достоинства поправок парадного входа не остались не оцененными. В 2014 году Адам Глинн и Константин Кашин, оба политологи из Гарварда (Глинн впоследствии перешел в Университет Эмори), написали получившую премию работу, которую следовало бы сделать обязательным чтением для всех ученых, занимающихся исчислениями в области общественных наук. Они применили новый метод к массиву данных, ранее тщательно изученных представителями общественных наук, — исследованию по Закону о партнерстве в области профессиональной подготовки (*Job Training Partnership Act; JTPA*), которое проводилось с 1987 по 1989 год. По результатам JTPA 1982 года, Департамент труда создал программу профессиональной подготовки, которая, помимо других целей, снабжала участников профессиональными навыками, навыками поиска работы и опытом работы. Она собирала данные о людях, подававших заявки для участия в этой программе, тех, кто реально пользовались ее услугами, и об их доходах за последующие 18 месяцев. Следует обратить внимание, что в исследование входили и РКИ, и данные, полученные в результате наблюдений, в которых люди делали выбор самостоятельно.

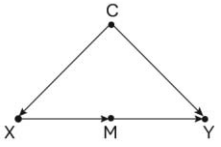


Рис. 42. Базовые условия для применения критерия парадного входа

Глинн и Кашин не рисовали каузальных диаграмм, но, судя по описанию их исследования, я бы нарисовал ее так, как на рис. 43. Переменная *записавшиеся* сообщает, зарегистрировался ли испытуемый для участия в программе или нет; переменная *посещавшие* сообщает, посещали ли записавшиеся занятия на самом деле. Очевидно, что программа могла повлиять на доходы только в том случае, если пользователь действительно посещал занятия, поэтому отсутствие прямой стрелки от записавшихся к доходам легко понять.

Глинн и Кашин не стали вдаваться в природу осложнителей, но я просуммировал их в переменной *мотивация*. Ясно, что человек, испытывающий сильную мотивацию увеличить свои доходы, с большей вероятностью запишется на курсы. Этот же человек с большей вероятностью увеличит свой заработок через 18 месяцев, вне зависимости от того, посещал ли он курсы. Цель исследования, конечно, — отделить влияние этого осложняющего фактора и найти, насколько велика помощь непосредственно от курсов.

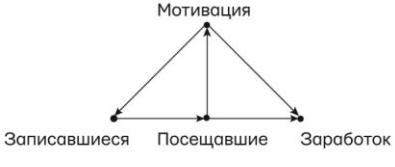


Рис. 43. Каузальная диаграмма для исследования JTPA

Сравнивая рис. 42 и 43, мы увидим, что критерий парадного входа был бы здесь применим, если бы не было стрелки

от *мотивации* к *посещавшим* — щита, упомянутого выше. Во многих случаях мы оправдываем отсутствие такой стрелки. Например, если бы услуги программы осуществлялись только в назначенное время и люди не укладывались в него только по уважительным причинам, не связанным с мотивацией (скажем, забастовка водителей общественного транспорта или сломанная нога), мы могли бы стереть эту стрелку и воспользоваться критерием парадного входа.

В реальных условиях исследования, когда услуги программы доступны в любое время, подобный аргумент не годится. Тем не менее — и это особенно интересно — Глинн и Кашин протестировали критерий парадного входа. Отнесемся к этому как к тесту на сенситивность. Если мы подозреваем, что средняя стрелка обозначает очень слабое воздействие, искажение, возникающее, если считать ее отсутствующей, совсем незначительно. Судя по их результатам, именно так дело и обстояло. Приняв определенные разумные допущения, Глинн и Кашин получили неравенства, по которым определили, была ли поправка чрезмерной или недостаточной и насколько. Наконец, они сравнили предсказания черного хода и парадного входа с результатами рандомизированного контролируемого исследования, которое проводилось в то же самое время. Результаты впечатлили. Оценки с помощью критерия черного хода (с поправками по таким известным конфаундерам, как возраст, раса и регион) оказались совершенно неверны, они отличались от экспериментальных результатов на сотни тысяч долларов. Это именно та картина, которая наблюдается, если имеется нераспознанный конфаундер. Критерий черного хода не способен внести по нему поправки. Тем не менее оценки парадного входа убрали почти все воздействия со стороны переменной *мотивация*. Для мужчин оценки по критерию парадного входа оказались в пределах экспериментальной ошибки РКИ, даже с небольшой положительной ошибкой, предсказанной Глинном и Кашиным. Для женщин результаты былиь еще точнее. Оценки парадного входа совпали с экспериментальными данными почти идеально, без сколько-нибудь заметной ошибки. Работа Глинна и Кашина подтверждает как

эмпирически, так и экспериментально, что, если только воздействие *C* на *M* (на рис. 42) незначительно, поправки парадного входа могут дать разумно точную оценку воздействия *X* на *Y*. Результат при этом значительно лучше, чем если вовсе не вводить поправок по *C*.

Изыскания Глинна и Кашина показывают, почему поправки парадного входа оказываются столь мощным инструментом: он позволяет нам снимать осложнения по таким переменным, по которым мы не можем получить наблюдений (например, в случае мотивации), включая те, которые даже не можем никак назвать. Рандомизированные контролируемые исследования считаются золотым стандартом оценок каузального воздействия ровно по тем же причинам. Поскольку оценки парадного входа равноценны, к тому же обладают дополнительным преимуществом, позволяя наблюдать поведение людей в их привычной обстановке, а не в условиях лаборатории, я не удивлюсь, если когда-нибудь этот метод составит серьезную конкуренцию РКИ.

Математика *Do*-оператора, или сознание над материей

Главная цель обеих обсужденных выше поправок — вычислить эффект интервенции, $P(Y \mid do(X))$, в терминах данных типа $P(Y \mid X, A, B, Z\ldots)$, не включающих оператор *do*. Если нам удастся полностью устранить все *do*, мы сможем использовать для оценки каузального воздействия наблюдаемые данные, что позволит нам перепрыгнуть со ступени 1 на ступень 2 Лестницы Причинности.

Тот факт, что в приведенных двух случаях мы это сделали (черный ход и парадный вход), немедленно поднимает вопрос, существуют ли другие входы и выходы, через которые устраняются все *do*. Рассуждая в общем и целом, мы поднимаем вопрос, реально ли решить заранее, допускает данная каузальная модель подобную процедуру устранения или нет. Если да, мы применим эту процедуру и обретем желаемое каузальное воздействие, не пошевелив пальцем для осуществления

интервенции. В противном случае мы по крайней мере будем знать, что допущения, встроенные в модель, недостаточны для того, чтобы выявить каузальное воздействие с помощью одних только наблюдений, и, как бы умны мы ни были, нам никуда не деться от постановки интервенционного эксперимента того или иного рода.

Перспектива принятия таких решений на основе чисто математических средств должна показаться заманчивой любому, кто понимает дороговизну и сложность проведения рандомизированных контролируемых исследований даже в тех случаях, когда они возможны с точки зрения физики и законодательства. Я тоже был вдохновлен этой идеей в начале 1990-х, не как экспериментатор, а как ученый в области информатики и заодно философ. Несомненно, одно из самых радостных событий в жизни ученого — обнаружить, что, не выходя из-за своего стола, вы способны определить, что возможно или невозможно в реальном мире — особенно, если решаемая проблема важна для общества, а тех, кто пытался ее решить до вас, она ставила в тупик. Могу себе представить, что нечто подобное испытывал Гиппарх из Никеи, когда обнаружил, что в состоянии вычислить высоту пирамиды по ее тени на земле, не взбираясь на нее. Это была явная победа разума над материей.

В самом деле, используемый мной подход во многом был вдохновлен учеными Древней Греции (включая Гиппарха) и изобретенной ими формальной логикой геометрии. В центре древнегреческой логики — набор аксиом, или самоочевидных истин, допустим: «Между двумя точками можно провести одну и только одну прямую». С помощью этих аксиом древним грекам удалось создать сложные структурированные утверждения — теоремы, истинность которых уже очень далека от очевидной. Возьмем, к примеру, утверждение, что сумма углов треугольника равна 180° (или двум прямым углам) вне зависимости от его размера и формы. Истинность этого утверждения ни в какой мере не очевидна; однако философы-пифагорейцы V века до н.э. сумели доказать его универсальную истинность, используя самоочевидные аксиомы в качестве деталей конструктора.

Если вы постараетесь вспомнить школьные уроки геометрии, хотя бы в первом приближении, вы вспомните, что доказательства теорем всегда состоят из вспомогательных построений: скажем, прямой, параллельной стороне треугольника, отмечающей равенство определенных углов; окружности с радиусом, равным данному сегменту, и т.д. Эти вспомогательные построения рассматриваются как временные математические предложения, которые содержат допущения (или требования), касающиеся свойств изображенных фигур. Каждое новое построение опирается на уже существующие, так же как и на аксиомы и на ранее доказанные теоремы. Например, начертание прямой, параллельной одной из сторон треугольника, определяется пятой аксиомой Евклида, о том, что возможно провести одну и только одну прямую, параллельную данной прямой через точку, не лежащую на этой прямой. Начертание этих вспомогательных конструкций — всего лишь операция механического манипулирования символами: в ходе него предложение, написанное ранее (или ранее начертанное изображение), переписывается в новом формате, если это переписывание допускается аксиомой. Великая заслуга Евклида в том, что он определил минимальный набор из всего пяти аксиом, из которого возможно вывести все остальные истинные утверждения геометрии.

Теперь давайте вернемся к нашему центральному вопросу: в каких случаях модель может заменить эксперимент или когда данные, полученные в результате действия, можно заменить просто наблюдаемыми данными. Вдохновившись геометрами Древней Греции, мы хотели бы свести задачу к манипуляции символами и таким образом свергнуть причинность с Олимпа и сделать ее доступной обычному исследователю.

Для начала перефразируем задачу нахождения воздействия X на Y, используя язык доказательств, аксиом и вспомогательных построений, язык Евклида и Пифагора. Начнем с нашего итогового предложения $P(Y \mid do(X))$. Задача будет считаться выполненной, если нам удастся удалить из него do-оператор, оставив только классические выражения для вероятностей, вроде $P(Y \mid X)$ или $P(Y \mid X, Z, W)$. Конечно, мы не вправе мани-

пулировать нашим итоговым, целевым выражением так, как нам вздумается; операции должны соответствовать тому, что *do* (*X*) означает как физическая интервенция. Таким образом, необходимо провести наше выражение через последовательность законных манипуляций, каждая из которых разрешена аксиомами и допущениями нашей модели. Эти манипуляции будут сохранять значение выражения, которое им подвергается, изменяя только формат, в котором оно записывается.

Пример преобразования, сохраняющего значение, — алгебраическое преобразование, превращающее $y = ax + b$ *в* $ax = y - b$. Отношения между *X* и *Y* остаются прежними, меняется только формат.

Мы уже знакомы с некоторыми легитимными преобразованиями *do*-выражений. Так, правило 1 гласит, что, когда мы наблюдаем переменную *W*, которая не имеет отношения к *Y* (возможно, является условной по отношению к другим переменным *Z*), вероятностное распределение *Y* не изменится. В главе 3 мы видели, что переменная пожар нерелевантна для состояния переменной тревога, если мы знаем состояние переменной-медиатора (дым). Это утверждение о нерелевантности переводится как символическая манипуляция: $P(Y \mid do(X), Z, W) = P(Y \mid do(X), Z)$. Постулированное выше уравнение правомерно, если набор переменных *Z* блокирует все пути от *W* к *Y* после того как мы удалили все стрелки, ведущие к X. В примере пожар → дым → тревога *W* = пожар, *Z* = дым и *Y* = тревога, а *Z* блокирует все пути от *W* к *Y* (в этом случае у нас нет переменной *X*).

Следующая легитимная трансформация знакома нам по обсуждению критерия черного хода. Мы знаем, что, если набор переменных *Z* блокирует все пути черного хода от *X* к *Y*, поправка по *Z*, $do(X)$ эквивалентна $see(X)$.

Следовательно, мы можем написать $P(Y \mid do(X), Z) = P(Y \mid X, Z)$, если *Z* удовлетворяет критериям черного хода. Примем это как правило 2 нашей системы аксиом. Хотя это, вероятно, менее самоочевидное правило, чем правило 1, в простейших случаях это принцип общей причины Ханса Рейхенбаха, измененный таким образом, чтобы мы не путали схождения

с конфаундерами. Другими словами, мы говорим, что после того, как введены поправки по достаточному набору переменных, снимающих осложнения, любая оставшаяся корреляция представляет собой истинное каузальное воздействие.

Правило 3 очень простое: оно более-менее сводится к тому, что мы можем убрать $do\ (X)$ из $P\ (Y \mid do\ (X))$ в любых случаях, в которых нет каузальных путей от X к Y, т.е. $P\ (Y \mid do\ (X)) = P\ (Y)$, если нет пути от X к Y, состоящего только из стрелок, направленных вперед. Перефразируем это правило следующим образом: если мы делаем нечто, что не влияет на Y, вероятностное распределение Y не изменяется. Помимо того, что правила 1—3 столь же самоочевидны, как и аксиомы Евклида, их можно также доказать математически, используя наше «бесстрелочное» определение *do*-оператора и базовые законы вероятности. Обратите внимание, что правила 1 и 2 включают условные вероятности, связанные со вспомогательными переменными Z, отличными от X и Y. Эти переменные допустимо считать контекстом, в котором исчисляется вероятность. Иногда уже само присутствие этого контекста делает преобразования законными. В правиле 3 также могут присутствовать вспомогательные переменные, но я опустил их для простоты.

Отмечу, что у каждого правила имеется простая синтаксическая интерпретация. Правило 1 разрешает добавить или удалить наблюдения. Правило 2 разрешает замену интервенции на наблюдение или наоборот. Правило 3 разрешает добавлять или удалять интервенции. Все эти разрешения действуют при определенных условиях, которые в каждом конкретном случае должны быть подтверждены каузальными диаграммами.

Теперь мы готовы продемонстрировать, как правила 1—3 позволяют нам преобразовывать одну формулу в другую до тех пор, пока (если только мы окажемся достаточно сообразительны) не получим выражение, которое нам нужно. Хотя это займет довольно много места, я думаю, что нужно все-таки наглядно показать вам, как с помощью последовательного применения правил *do*-исчисления получается формула парадного входа (рис. 44). Вам нет необходимости внимательно

следить за каждым шагом, я показываю вам вывод формулы, чтобы вы ощутили вкус *do*-исчисления.

Наше путешествие начнется с целевого выражения $P\ (Y \mid do\ (X))$. Мы вводим вспомогательные переменные и трансформируем целевое выражение так, чтобы оно не содержало оператора *do* и совпадало, конечно, с формулой поправок парадного входа. Каждый наш шаг обосновывается каузальной диаграммой, связывающей *X*, *Y* и вспомогательные переменные, или в некоторых случаях субдиаграммами, в которых стерты стрелки, соответствующие интервенциям. Эти обоснования изображаются справа.

К *do*-исчислению я испытываю особые чувства. С помощью этих трех скромных правил мне удалось вывести формулу парадного входа. Это было первое каузальное воздействие, которое получилось оценить иными средствами, чем поправки по конфаундерам. Я был убежден, что без *do*-исчисления этого никто не сможет сделать, поэтому представил эту задачу как вызов на семинаре по статистике в Калифорнийском университете в Беркли в 1993 году и даже предложил приз в 100 долларов тому, кто ее решит. Пол Холланд, присутствовавший на семинаре, написал мне, что предложил задачу в качестве проекта своим студентам и пошлет мне решение, когда оно будет готово (коллеги рассказывали мне, что на конференции в 1995 году он таки представил длинное решение, так что, возможно, я должен ему 100 долларов, если найду его доказательство). Экономисты Джеймс Хекман и Родриго Пинто предприняли следующую попытку доказать формулу парадного входа, используя только «стандартные инструменты», в 2015 году. Им это удалось, хотя и ценой восьми страниц сложных выкладок.

В ресторане вечером накануне этой беседы я записал доказательство (очень похожее на то, что приведено на рис. 44) на салфетке для Дэвида Фридмана. Позже он написал мне, что потерял ту салфетку, не может восстановить доказательство, и спросил меня, не сохранилось ли у меня копии. На следующий день Джейми Робинс написал мне из Гарварда, сообщив, что слышал о «задаче на салфетке» от Фридмана и готов вылететь

в Калифорнию ближайшим рейсом, чтобы вывести доказательство вместе со мной.

do-исчисление в действии

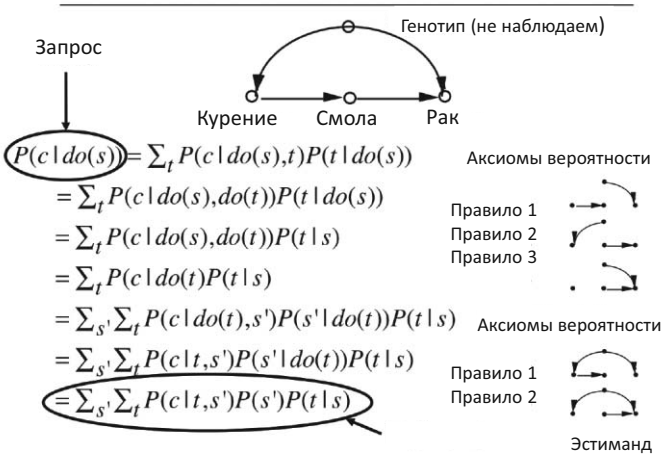


Рис. 44. Вывод формулы поправок парадного входа по правилам *do*-исчисления

Я был очень рад поделиться с Робинсом секретами *do*-исчисления и полагаю, что его поездка в Лос-Анджелес в том году сыграла ключевую роль в том, с каким энтузиазмом он воспринял каузальные диаграммы. Благодаря ему и Сандеру Гренланду эти диаграммы стали вторым языком эпидемиологов. Это объясняет, почему я так отношусь к «задаче на салфетке».

Поправка парадного входа была приятным сюрпризом и означала, что *do*-исчислению есть что предложить людям. Тем не менее в тот момент я еще не знал наверняка, достаточно ли всего трех правил *do*-исчисления. Не упустили ли мы четвертое правило, которое помогло бы нам решать задачи, неразрешимые с помощью только этих трех?

В 1994 году, когда я впервые предложил общественности *do*-исчисление, я выбрал эти три правила потому, что они были достаточны во всех известных мне случаях. Я не знал,

выведут ли они меня как нить Ариадны из абсолютно любого лабиринта, или когда-нибудь мне попадется лабиринт такой дьявольской сложности, что выбраться из него я не смогу. Конечно, я надеялся на лучшее. Я предполагал, что в тех случаях, когда каузальное воздействие вообще возможно оценить по данному набору данных, последовательность шагов, использующих эти три правила, позволит сократить *do*-оператор. Но я не в состоянии был это доказать.

У такого типа задач есть множество предшественников в математике и логике. Это свойство в математической логике обычно называют функциональной полнотой. У обладающей полнотой системы аксиом есть следующее свойство: этих аксиом достаточно для выведения любого истинного утверждения данного языка. Некоторые очень хорошие системы аксиом тем не менее не обладают функциональной полнотой: таковы, например, аксиомы Филипа Давида, описывающие условную независимость в теории вероятности.

В этом современном мифе о лабиринте роль Ариадны для моего блуждающего Тезея сыграли две группы исследователей: Имин Хуан и Марко Вальторта из Университета Южной Каролины, а также мой собственный студент Илья Шпицер из Калифорнийского университета в Лос-Анджелесе. Обе группы одновременно и независимо доказали, что правил 1—3 достаточно, для того чтобы выбраться из любого лабиринта, из которого в принципе есть выход. Я не уверен в том, что весь мир, затаив дыхание, ждал доказательства функциональной полноты этих аксиом, потому что большинству исследователей в то время хватало просто критериев черного хода и парадного входа. Тем не менее обе команды получили награды в номинации лучших студенческих работ на конференции «Неопределенность в искусственном интеллекте», проходившей в 2006 году.

Однако признаюсь, что я как раз ждал этого решения, затаив дыхание. Оно сообщает нам, что если мы не в состоянии найти способ оценить вероятность $P(Y \mid do(X))$ с помощью правил 1—3, то решения не существует. В этом случае мы знаем, что без рандомизированного контролируемого исследования не обойтись. Оно также говорит нам, какие дополнительные

допущения или эксперименты нужны для того, чтобы каузальное воздействие стало возможно оценить.

Перед тем как объявить окончательную победу, нам нужно обсудить одну проблему с *do*-исчислением. Как и любые вычислительные методы, оно помогает нам выстроить доказательство, но не найти решение. С ним замечательно легко проверить истинность решения, но для поиска решений оно не так хорошо. Если вы знаете правильную последовательность преобразований, легко продемонстрировать другим (знакомым с правилами 1—3), что *do*-оператор можно сократить. Однако, если правильная последовательность вам неизвестна, не так-то легко обнаружить ее или даже определить, что она вообще существует. Используя аналогию с геометрическими доказательствами, нам надо решить, какие дополнительные построения потребуются нам на следующем шаге. Окружность с центром в точке *A*? Прямая, параллельная *AB*? Число возможностей безгранично, а аксиомы сами по себе не дают намека, что делать дальше. Мой школьный учитель геометрии говорил, что нужно посмотреть на ситуацию через «математические очки».

В математической логике это называется проблемой принятия решений. Многие логические системы страдают от неустранимой проблемы принятия решений. Например, если у нас есть кучка костяшек домино различных размеров, у нас нет определенного способа решить, возможно ли сложить из них квадрат заданного размера. Но когда расклад уже заявлен, очень просто подтвердить, является ли он решением.

К счастью (опять) для *do*-исчисления, с проблемой принятия решений удалось справиться. Илья Шпицер, основываясь на более ранней работе другого моего студента по имени Цзинь Тянь, нашел алгоритм, который определяет, существует ли решение в «полиномиальном времени». Это до некоторой степени технический термин, но, продолжая нашу аналогию с поиском выхода из лабиринта, это означает, что у нас появляется намного более эффективный способ поиска выхода, чем случайное блуждание по всем доступным путям.

Алгоритм Шпицера для поиска всех каузальных воздействий не устраняет потребности в *do*-исчислении. На самом деле мы

нуждаемся в нем даже больше, и по нескольким независимым причинам. В первую очередь, оно нужно нам для того, чтобы продвинуться дальше исследований, использующих только наблюдения. Допустим, что все складывается наихудшим образом и наша каузальная модель не допускает оценки каузального воздействия $P(Y \mid do(X))$ исключительно из данных, полученных в результате наблюдений. Предположим, что мы также не в состоянии провести рандомизированное контролируемое исследование со случайно назначаемыми X. Сообразительный исследователь спросит, удастся ли нам тогда оценить $P(Y \mid do(X))$, рандомизируя какую-либо другую переменную, скажем Z, которая лучше поддается контролю, чем X.

Например, если мы хотим узнать воздействие уровня холестерина (X) на сердечно-сосудистые заболевания (Y), мы можем манипулировать диетой испытуемых (Z) вместо того, чтобы пытаться прямо контролировать уровень холестерина у них в крови. Таким образом, мы задаемся вопросом, реально ли найти такую суррогатную переменную Z, которая позволит получить ответ на вопрос о причинах. На языке *do*-исчисления вопрос звучит так: возможно ли найти такое Z, чтобы преобразовать $P(Y \mid do(X))$ в выражение, в котором Z, а не X подвергается *do*-оператору. Это совершенно отдельная задача, не подпадающая под алгоритм Шпицера. К счастью, для нее тоже есть полный ответ — новый алгоритм, открытый Элиасом Баренбоймом в моей лаборатории в 2012 году.

Еще больше подобных задач возникает, когда мы рассматриваем проблемы транспортабельности или внешней валидности данных, выясняя, будет ли экспериментальный результат сохранять валидность в других условиях, отличающихся от изученных по нескольким ключевым параметрам. Этот более амбициозный набор вопросов бьет уже в самое сердце научной методологии, потому что наука не существует без обобщений. Однако вопрос обобщения не двигался с места по крайней мере два столетия. Инструментов для нахождения решения попросту не было.

В 2015 году Баренбойм и я представили в Национальную академию наук публикацию, в которой проблема решается

в случае, если вы можете выразить свои допущения относительно двух сред с помощью каузальных диаграмм. В этих случаях правила *do*-исчисления обеспечивают систематический подход к определению того, насколько каузальные воздействия, определенные в опытных условиях, помогут нам оценить воздействия в интересующих нас целевых условиях.

Еще одна причина, по которой *do*-исчисление остается необходимым, — это прозрачность. Когда я писал эту главу, Баренбойм (ныне профессор в Университете Пердью) прислал мне новую головоломку — диаграмму, в которой всего четыре наблюдаемые переменные: *X*, *Y*, *Z* и *W* — и две ненаблюдаемые: *U*1 и *U*2 (рис. 45). Он предложил мне определить, возможно ли оценить воздействие *X* на *Y* в этом случае. Здесь невозможно заблокировать пути черного хода и не выполняются условия для парадного входа. Я испробовал все свои любимые подходы и ранее выручавшие меня интуитивные аргументы, как за, так и против, и все же я не мог понять, как это сделать. Я не мог найти выхода из лабиринта. Но как только Баренбойм шепнул мне: «Попробуй *do*-исчисление!», ответ немедленно засиял во всей непосредственной простоте. Каждый шаг стал ясен и наполнен смыслом. Теперь это самая простая из известных нам моделей, в котором каузальное воздействие приходится оценивать методом, выходящим за рамки поправок черного хода и парадного входа.

Чтобы не оставлять читателя под впечатлением, что *do*-исчисление хорошо только для теории и для решения головоломок на досуге, я закончу эту секцию практической задачей, которую поставили недавно два ведущих статистика Нэнни Вермут и Дэвид Кокс. Она показывает, как дружеская подсказка «Попробуй *do*-исчисление» может помочь экспертам в области статистики решать сложнейшие практические задачи.

Приблизительно в 2005 году Вермут и Кокс заинтересовались так называемой проблемой последовательных решений, или курса лечения, меняющегося во времени, с которой часто приходится сталкиваться, например, при лечении СПИДа. В этом случае типично, что лечение проводится продолжительное время, и в каждый период времени врачи изменяют

силу и дозировку медикаментозных средств в последующих назначениях, исходя из состояния здоровья пациента. В свою очередь, на состояние здоровья пациента влияет лечение, назначенное ему в прошлом.

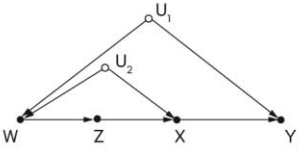


Рис. 45. Новая «задача на салфетке»?

Таким образом, в итоге у нас получается сценарий вроде того, что изображен на рис. 46, где показаны два отрезка времени и два назначения врача. Первое назначение (*X*) рандомизировано, а второе (*Z*) назначается в ответ на наблюдение (*W*), которое зависит от *X*. Задача Кокса и Вермут состояла в том, чтобы из данных, полученных в таком режиме лечения, предсказать воздействие X на исход *Y*, предполагая, что *Z* остается постоянным во времени вне зависимости от наблюдений *W*.

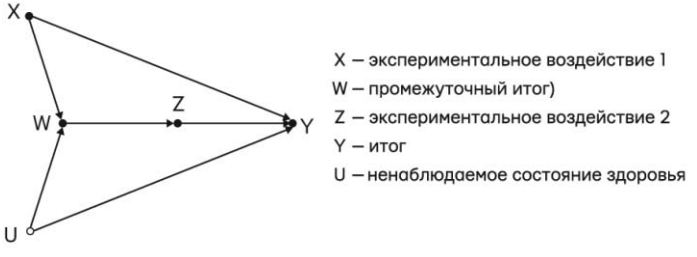


Рис. 46. Пример меняющегося во времени курса лечения, по Вермут и Коксу

Джейми Робинс впервые обратил мое внимание на проблему меняющегося во времени курса лечения в 1994 году, и с помощью *do*-исчисления мы получили общее решение, включающее серийную версию формулы поправок черного хода. Вермут и Кокс, не зная об этом методе, назвали свою

проблему «непрямое осложнение» и опубликовали три работы, посвященные ее анализу (2008, 2014 и 2015). Не сумев решить ее в общем виде, они прибегли к линейной аппроксимации, но даже в линейном виде решение показалось им сложным и неудобным, потому что стандартными методами регрессии решить эту задачу нельзя.

К счастью, когда муза шепнула мне в ухо «Попробуй *do*-исчисление», я заметил, что их задача решается в три строчки вычислений. Логика решения такова. Нам нужно вычислить $P(Y \mid do(X), do(Z))$, в то время как нам доступны данные в виде $P(Y \mid do(X), Z, W)$ и $P(W \mid do(X))$. Это отражает тот факт, что в исследовании, откуда получены наши данные, Z не контролируется извне, а следует за W согласно некоему (неизвестному) протоколу. Таким образом, наша задача — преобразовать искомое выражение в другое выражение, отражающее условия исследования, при которых *do*-оператор применяется только к X, но не к Z. Оказывается, что одно только применение трех правил *do*-исчисления позволяет этого добиться. В этой истории нет никакой морали, кроме глубокого почтения к возможностям математики по решению сложнейших задач, иногда влекущих за собой практические последствия.

Гобелен, сотканный наукой, или невидимые музыканты *Do*-оркестра

Я уже упоминал роль некоторых моих студентов в создании великолепного гобелена *do*-исчисления. Как и любой гобелен, со стороны он производит впечатление цельности и законченности, за которыми не видно, какими трудами он был создан и как много рук участвовало в процессе. В этом случае потребовалось более двух десятилетий и участие нескольких моих студентов и коллег.

Первым был Томас Верма, с которым я познакомился, когда он был еще подростком 16 лет. Его отец привел его однажды в мой офис и сказал: «Займите его чем-нибудь». Он был слишком талантлив и никому из преподавателей математики в высшей

школе не удавалось найти что-нибудь достаточно интересное для него. То, чего он добился, по-настоящему потрясало воображение. Верма в итоге доказал то, что стало впоследствии известно как свойство *d*-сепарации (т.е. тот факт, что можно использовать правила блокирования путей для определения, какие независимые наборы переменных должны сохраняться в данных). Поразительно, но он сообщил мне, что доказал свойство d-сепарации, считая, что это была уже давно решенная задача, которую задают в качестве домашнего задания для студентов! Иногда быть молодым и наивным оказывается полезно.

Он внес свой вклад в правило 1 *do*-исчисления и в идею, что перекрывание путей остается на первой ступени Лестницы Причинности.

Сила доказательства Верма не была бы настолько высоко оценена, если бы не комплементарный ему результат, показывающий, что его нельзя улучшить. Другими словами, каузальная диаграмма не подразумевает никакие другие наборы независимых переменных, кроме тех, которые выявляются блокированием путей. Этот этап закончил другой студент, Дэн Гейгер. Он перешел в мою лабораторию из другой исследовательской группы в Калифорнийском университете в Лос-Анджелесе, после того как я пообещал ему, что он сразу получит степень кандидата наук, если сможет доказать две теоремы. Он выполнил свою часть обещания — выполнил свою и я! Теперь он декан отделения компьютерных наук в израильском Технионе[1], моей альма-матер.

Но Дэн был не единственным студентом, кого мне удалось переманить с другого факультета. Однажды в 1997 году, одеваясь в раздевалке бассейна университета, я разговорился с парнем-китайцем по соседству. Он писал кандидатскую по физике, и, как было тогда в моих обычаях, я агитировал его переключиться на искусственный интеллект, где происходит вся «движуха». Убедить его до конца не удалось, но уже

[1] Израильский технологический институт *(прим. ред)*.

на следующий день мне пришло письмо на электронную почту от его друга Цзинь Тяня, в котором тот сообщал, что хотел бы переключиться с физики на науку о компьютерах, и спрашивал, не найдется ли у меня для него интересного проекта на лето? Через два дня он уже работал у меня в лаборатории.

Через четыре года, в апреле 2001-го, он удивил мир, обнаружив простой графический критерий, обобщающий черный ход, парадный вход и все прочие ходы, которые мы только могли себе тогда вообразить. Я помню, как впервые представил критерий Тяня на конференции в Санта-Фе. Один за другим ведущие исследователи смотрели на мой постер и качали головами, не веря своим глазам. Как такой простой критерий может быть применим к любым диаграммам?

Тянь (в настоящее время профессор государственного Университета Айовы) пришел в нашу лабораторию со своим стилем мышления, который нам тогда, в 1990-е, казался странным и чужим. Наши беседы были набиты дикими метафорами и совершенно сырыми гипотезами. Но Тянь никогда не произносил ни одной фразы, в которой не был бы абсолютно, железобетонно уверен, любые его слова подтверждались доказательством. Последующее слияние двух стилей принесло свои плоды. Метод Тяня, называемый *c*-декомпозицией, позволил Илье Шпицеру разработать полный алгоритм *do*-исчисления. Мораль: не стоит недооценивать значение разговоров в спортивной раздевалке!

Илья Шпицер появился у нас под конец нашей на тот момент уже десятилетней борьбы с пониманием интервенций. Его появление совпало с очень сложным периодом, когда я вынужден был отложить дела, занимаясь созданием фонда в память о моем сыне Даниэле, который стал жертвой терроризма. Я всегда жду от своих студентов, что они будут самостоятельными и самодостаточными, но к тем, кто учился у меня тогда, это правило оказалось применено в крайней степени. Они сделали мне лучший из возможных подарков, положив последние, но принципиально важные стежки на гобелен *do*-исчисления, чего я не был в состоянии сделать сам. На самом деле я даже пытался отговорить Илью от попыток доказать функциональную полноту *do*-исчисления. Доказательства полноты особенно

сложны, и студентам, надеющимся закончить аспирантуру вовремя, таких кандидатских тем лучше избегать. К счастью, Илья сделал все сам у меня за спиной.

Коллег также следует благодарить за огромное влияние на ваше мышление в критические моменты. Питер Спертс, профессор философии в Университете Карнеги — Меллона, раньше меня пришел к сетевому подходу к причинности, и его влияние на меня было основополагающим. На его лекции в Упсале я впервые узнал, что осуществление интервенций может быть представлено как стирание стрелок в каузальных диаграммах. До этого я работал под тем же тяжким ярмом, что и многие поколения статистиков, пытающихся рассматривать причинность в терминах «одна диаграмма — одно статическое вероятностное распределение».

Идея стирания стрелок не принадлежала исключительно Спертсу. В 1960 году экономисты Роберт Стротц и Герман Вольд предложили практически ту же идею. Тогда в экономике диаграммы использовать было не принято; вместо них экономисты полагались на модели структурных уравнений, которые представляют собой, по сути, то же, что и уравнения Сьюалла Райта, только без диаграмм. Удалению стрелки в путевой диаграмме соответствует удаление уравнения из модели структурных уравнений. Таким образом, в более общем смысле Стротц и Вольд пришли к этой идее первыми, если мы, конечно, не хотим зарываться еще глубже в историю науки (им предшествовал Трюгве Ховельмо, норвежский экономист и лауреат Нобелевской премии), который в 1943 году предложил модифицировать уравнения для отражения интервенций. Тем не менее предложенный Спертсом перевод удаления уравнений в термины каузальных диаграмм вызвал лавину новых идей и достижений. Критерий черного хода первым выиграл от перехода на новый язык, следующим шло *do*-исчисление.

Однако эта лавина все еще не остановилась. Прогресс в таких сферах, как контрфактивность, обобщаемость, недостающие данные и машинное обучение, продолжает бурлить.

Если бы я был менее скромен, я бы закончил этим, приведя напоследок известное высказывание Исаака Ньютона про «стоя-

ние на плечах гигантов». Но, оставаясь собой, я не могу устоять перед искушением вместо этого процитировать Мишну: «Харбе ламадети миработай ум'хаверай йотер мехем, умиталмидай йотер микулам», что означает: «Многому я научился у своих наставников, еще более — у своих товарищей, но более всего — у своих учеников» (Таанит, 7а). *Do*-оператор и *do*-исчисление не существовали бы сегодня в современном виде без весомых вкладов Верма, Гейгера, Тяня, Шпитцера и др.

Любопытная история с доктором Сноу

В 1853—1854 годах Англию терзала эпидемия холеры. В те времена холера была страшнее, чем сегодня Эбола: здоровый человек, выпив зараженную холерой воду, может умереть уже через сутки. Сегодня нам известно, что холеру вызывает бактерия — холерный вибрион, поражающий кишечник. Распространяясь, он вызывает у своих жертв неудержимый понос типа «рисовый отвар»; потеряв с диареей огромное количество жидкости, больной умирает.

Однако в 1853 году еще никто не видел под микроскопом ни одной бактерии, не говоря уже о возбудителе холеры. Тогда считалось, что холеру вызывает нездоровый воздух, «миазмы», и эта теория на первый взгляд подтверждалась тем, что от эпидемии гораздо сильнее страдали самые бедные районы Лондона, где царила антисанитария. Доктор Джон Сноу, врач, занимавшийся жертвами холеры более 20 лет, к теории миазмов всегда относился скептически. Он разумно рассуждал, что, если симптомы проявляются в первую очередь в желудочно-кишечном тракте, заболевание должно вызываться попаданием вызывающего его агента в кишечник. Но поскольку возбудителя заболевания нельзя было увидеть, он не мог это доказать — до эпидемии 1854 года.

В истории Джона Сноу две главы, и одна намного более известна, чем другая. В первой, которую можно назвать голливудской версией, он, рискуя жизнью, ходит из дома в дом, выясняя, где люди умерли от холеры, и обнаруживает огром-

ную концентрацию смертей, с десятками погибших, вокруг колодца с насосом на Броуд-стрит. Разговаривая с жителями этого района, он выясняет, что практически все погибшие брали воду из этого конкретного колодца. Ему даже становится известно о смерти, случившейся достаточно далеко от этого места, в Хэмпстеде. Одной женщине оттуда понравился вкус воды из колодца на Броуд-стрит, и она вместе со своей племянницей пила воду именно оттуда. Обе они умерли, хотя никто в ее районе даже не заболел. Собрав воедино все эти факты, Сноу требует от местных властей убрать рукоятку насоса, чтобы прекратить забор воды, и 8 сентября власти соглашаются. Как пишет биограф Сноу, «ручку насоса сняли, и моровое поветрие удалось остановить».

Эта история замечательно кинематографична. В наше время общество имени Джона Сноу даже проводит торжественную театральную постановку, изображающую снятие ручки насоса у колодца. Однако, если смотреть правде в глаза, закрытие колодца вряд ли было заметно на фоне общегородской эпидемии, от которой по-прежнему умирало почти 3 тысячи человек (в день?).

В другой, уже не голливудской, серии этого фильма, мы снова видим доктора Сноу, пешком обходящего весь старый Лондон, но на этот раз он пытается выяснить, где все жители его города берут воду. В то время водопроводные услуги лондонцам предоставляли в основном две частные компании: «Саутворк и Воксхол» и «Ламбет». Как удалось выяснить Сноу, основное различие между ними было в том, что первая компания осуществляла забор воды из Темзы у Лондонского моста, ниже слива городской канализации. Вторая несколькими годами раньше переместила водозабор выше по течению, до канализационного слива. Таким образом, клиенты «Саутворка» получали воду, загрязненную канализационными стоками, а клиенты «Ламбета» — относительно чистую (обратите внимание, что оба этих водопровода не имели отношения к заразной воде с Броуд-стрит, которую брали из находящегося там отдельного колодца).

Статистика смертей легла в основу невеселой гипотезы Сноу. Кварталы, снабжавшиеся компанией «Саутворк и Воксхолл», особенно сильно страдали от холеры, и смертность в них была в восемь раз выше. Однако, несмотря на это, прямых доказательств под рукой не было. Защитники теории миазмов заявили бы, что ядовитые испарения были гораздо сильнее именно в этих районах, и их невозможно было бы опровергнуть. На языке каузальных диаграмм наша ситуация описывается рис. 47. Мы не в состоянии получить данные по конфаундеру *миазмы* (или другим конфаундерам, таким как *бедность*), поэтому мы не вправе ввести по нему поправки по методу черного хода.

Здесь Сноу додумался до поистине блестящей идеи. Он обнаружил, что в тех районах, куда был проведен водопровод из обеих компаний, смертность была все-таки значительно выше в домохозяйствах, получавших воду от «Саутворка». Однако они не отличались от соседних ни по уровню миазмов, ни по уровню бедности. «Водопроводы двух поставщиков переплетаются самым тесным образом, — писал Сноу. — Трубы каждой из двух компаний тянутся по каждой улице и входят почти в каждый двор и переулок. … Обе компании снабжают водой и богатых, и бедных, и большие дома, и маленькие домики: невозможно обнаружить разницы ни в благосостоянии, ни в роде занятий между гражданами, получающими воду от той или иной компании».

Хотя понятие о рандомизированном контролируемом исследовании было еще делом будущего, все выглядело так, будто водопроводные компании поставили на лондонцах РКИ. На самом деле Сноу даже обращает на это внимание: «Невозможно было бы спланировать опыт, который бы лучше выявил воздействие источника воды на распространение холеры, нежели этот, который обстоятельства в готовом виде предоставили наблюдателю. Размах этого опыта так же роскошен: не менее 300 тысяч людей обоих полов, всех возрастов и родов занятий, вне зависимости от чина и благосостояния, от дворянства до нищей бедноты, разделили на две группы без их спроса и в большей части случаев без их ведома».

Одна группа получала чистую воду; другая получала воду, загрязненную канализационными стоками.

Рис. 47. Каузальная диаграмма для холеры (до открытия холерного вибриона)

Наблюдения Сноу добавили к каузальной диаграмме еще одну переменную, и теперь она выглядит как рис. 48. Рискованное детективное исследование доктора Сноу привело к двум важным открытиям: 1) нет стрелки между миазмами и водопроводной компанией (эти две переменные независимы) и 2) есть стрелка между водопроводной компанией и чистотой воды. Третье обстоятельство не было упомянуто доктором Сноу, но не менее важно: 3) отсутствие прямой стрелки от водопроводной компании к холере, что сегодня для нас вполне очевидно, потому что теперь мы знаем, что водопроводные компании не доставляли холеру в дома своих клиентов каким-либо другим путем.

Переменная, которая удовлетворяет таким трем условиям, сегодня называется инструментальной переменной. Совершенно ясно, что Сноу воспринимал эту переменную как подбрасывание монеты, которое симулирует переменную без входящих стрелок. Поскольку во взаимоотношениях между переменными *водопроводная компания* и *холера* нет конфаундеров, любая наблюдаемая между ними связь должна быть причинно-следственной. Аналогично, поскольку воздействие водопроводной компании на холеру осуществляется через чистоту воды, мы (как и когда-то Сноу) заключаем, что наблюдаемая ассоциация между чистотой воды и холерой тоже должна быть причинно-следственной. Свой вывод Сноу вынес в недвусмысленных

терминах: если компания «Саутворк и Воксхол» перенесет водозабор выше по течению, это спасет тысячи жизней.

В то время на выводы доктора Сноу обратили внимание лишь немногие. Свои результаты он опубликовал в брошюре, изданной за его собственный счет: по рукам разошлись только 56 экземпляров этой брошюры. В наше время эпидемиологи рассматривают ее как основополагающий документ для всей своей дисциплины. Она показала, что старомодное расследование «на подметках ботинок» (эту фразу я позаимствовал у Дэвида Фридмана) вместе с применением каузальных рассуждений позволяют вычислить убийцу.

Хотя теория миазмов в наше время полностью развенчана, бедность в этом примере, несомненно, являлась конфаундером, как, впрочем, и местоположение. Однако, даже не собирая данные по этим переменным (так далеко опросы доктора Сноу не заходили), а используя лишь инструментальную переменную, мы способны вычислить, сколько жизней было бы спасено благодаря чистой воде.

Рис. 48. Диаграмма для холеры после введения инструментальной переменной

Вот как это работает. Для простоты мы вернемся к именам Z, X, Y и U для наших переменных и перерисуем диаграмму рис. 48 так, как на рис. 49. Я добавил путевые коэффициенты (a, b, c, d), отражающие силу каузальных воздействий. Мы, таким образом, предполагаем, что наши переменные исчислимы, а функции, описывающие их, линейны. Вспомним, что путевой коэффициент a означает, что интервенция по увеличению Z на одну стандартную единицу увеличит X на a стандартных единиц (здесь я опущу технические подробности о том, что такое «стандартная единица»).

Поскольку Z и X ничем не осложнены, каузальное воздействие Z на X (т.е. a) можно оценить по наклону r_{XZ} линии регрессии X на Z. Аналогично переменные Z и Y не осложнены, потому что путь $Z \rightarrow X \leftarrow U \rightarrow Y$ блокируется схождением по X. Следовательно, наклон линии регрессии Z на Y (r_{ZY}) будет равен каузальному воздействию на прямом пути $Z \rightarrow X \rightarrow Y$, которое представляет собой произведение путевых коэффициентов: ab. Итак, получаем два уравнения: $ab = r_{ZY}$ и $a = r_{ZX}$. Если мы разделим первое уравнение на второе, то получим каузальное воздействие X на Y: $b = r_{ZY} / r_{ZX}$.

Вот так инструментальные переменные позволяют выполнить тот же волшебный фокус, который нам удавался с помощью поправок парадного входа: мы нашли воздействие X на Y, даже не будучи в состоянии контролировать осложнитель U или получить по нему данные.

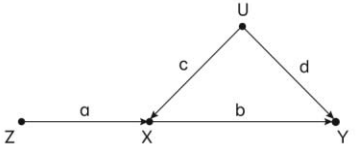


Рис. 49. Общая схема для инструментальных переменных

В итоге мы предоставили лицам, принимающим решения, убедительные аргументы о том, что водозабор надо передвинуть вверх по течению, даже если лица, принимающие решения, продолжают верить в теорию миазмов. Обратите также внимание, что мы добыли информацию со второго уровня Лестницы Причинности (b) из информации с первого уровня (корреляции r_{ZY} и r_{ZX}). Мы смогли это сделать, потому что допущения, воплощенные в путевой диаграмме, каузальны по своей природе, особенно критически важное допущение, что между переменными U и Z стрелки нет. Если бы каузальная диаграмма была иной, например если бы Z была конфаундером связи X и Y, формула $b = r_{ZY} / r_{ZX}$ не давала бы верной оценки

воздействия X на *Y*. На самом деле эти две модели невозможно различить никакими чисто статистическими методами, как бы велики ни были массивы данных.

Инструментальные переменные были известны до Революции Причинности, но каузальные диаграммы привнесли новую ясность в то, как они работают. Сноу воспользовался инструментальной переменной имплицитно, хотя у него и не было количественной формулы. Сьюалл Райт, несомненно, понимал пользу путевых диаграмм в этом случае; формула $b = r_{ZY} / r_{ZX}$ может быть напрямую выведена из его метода путевых коэффициентов.

Похоже, что первым ученым помимо самого Сьюалла Райта, кто сознательно воспользовался инструментальными переменными, был не кто иной, как… его собственный отец Филип Райт!

Вспомним, что Филип Райт был экономистом, работавшим в организации, которая впоследствии станет Брукингским институтом. Его интересовало, как объем производства некоего товара изменится, если будет введена пошлина, которая поднимет на товар цену и предположительно увеличит объем производства. На экономическом языке его интересовала эластичность предложения.

В 1928 году Райт написал объемную монографию, посвященную подсчетам эластичности предложения для льняного масла. В замечательном приложении к этой работе он анализирует вопрос с помощью путевых диаграмм. Это был смелый ход: вспомним, что ни один экономист в мире тогда не видел и не слышал ничего подобного (на самом деле Райт-старший подстраховался и подтвердил свои подсчеты также и с помощью более традиционных методов).

На рис. 50 показана несколько упрощенная версия диаграммы Райта. В отличие от большинства диаграмм в книге, которую вы читаете, на этой имеются стрелки, идущие в обе стороны, но я бы советовал читателю не слишком нервничать по этому поводу. С помощью некоторых математических трюков мы легко сможем заменить цепь *спрос* → *цена* → *предложение* одной стрелкой *спрос* → *предложение*, и диаграмма тогда

станет выглядеть как рис. 49 (хотя экономистам она будет после этого казаться менее приемлемой). Важно отметить, что Филип Райт осознанно ввел переменную *урожай (льняного масла) с акра* в качестве инструментальной, прямо влияющей на предложение, но не коррелирующей со спросом. Затем он применил метод анализа, подобный тому, который я только что привел, чтобы вычислить как влияние предложения на цену, так и влияние цены на предложение.

Историки науки спорят между собой о том, кто изобрел инструментальные переменные — метод, который стал очень популярен в современной эконометрике. Я не сомневаюсь в том, что Филип Райт позаимствовал идею путевых коэффициентов у своего сына. До этого ни один экономист не настаивал на различии между каузальными коэффициентами и коэффициентами регрессии, все они были в лагере Карла Пирса и Генри Найлза и считали, что причинность — это не более чем частный случай корреляции. Кроме этого, никто до Сьюалла Райта не предлагал способа вычисления коэффициентов регрессии на языке путевых коэффициентов с последующим выворачиванием процесса наизнанку, чтобы получить каузальные коэффициенты из регрессии. Это изобретение принадлежало исключительно Сьюаллу.

Вполне естественно, что некоторые историки экономики предположили, что все математическое приложение к этой монографии было написано Сьюаллом. Однако анализ стиля показал, что на самом деле автором был Филип.

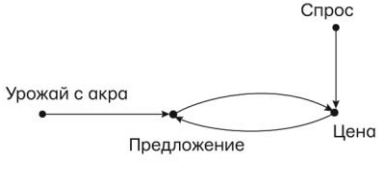


Рис. 50. Упрощенная версия каузальной диаграммы «предложение — цена» Райта-старшего

С моей точки зрения, такое детективное расследование дополнительно украшает эту историю. Оно показывает, что

Филип взял на себя труд разобраться в теории сына и выразить ее уже на своем языке.

Теперь переместимся из 50-х годов *XIX* века и 20-х XX-го в наши дни и посмотрим на инструментальные переменные за работой на примере, который я выбрал случайно из буквально многих десятков подобных.

«Хороший» и «плохой» холестерин

Помните ли вы момент, когда ваш семейный доктор впервые начал употреблять выражения «плохой» и «хороший» холестерин? Это могло случиться в 1990-е, когда препараты, снижающие уровень «плохого» холестерина (липопротеина низкой плотности, ЛПНП) в крови впервые появились на фармацевтическом рынке. Эти вещества, называемые статинами, вскоре стали приносить фармацевтическим компаниям доходы во многие миллиарды долларов.

Первым модифицирующим холестерин препаратом, испытанным с помощью рандомизированного контролируемого исследования, был холестирамин. Программа по предотвращению сердечно-сосудистых заболеваний, начатая в 1973 году и завершенная в 1984 году, показала сокращение холестерина в крови у мужчин, принимавших холестирамин, на 12,6% и 19%-ное снижение риска инфаркта. Поскольку речь идет об РКИ, вы можете подумать, что методы, описанные в этой главе, здесь не понадобятся, потому что они специально созданы для того, чтобы заменять РКИ в ситуациях, когда доступны только данные, полученные в результате наблюдений. Но это не так. В этом исследовании, как нередко случается с РКИ, проводимыми на людях, экспериментаторы столкнулись с проблемой неподчинения: испытуемые, которым случайным образом был назначен лекарственный препарат, на самом деле его не принимали. Это сокращает видимую эффективность препарата, поэтому логично попытаться ввести поправки по непослушным испытуемым. Но, как обычно, здесь тут же обнаруживается страшная личина конфаундеров. Если неподчинившиеся ис-

пытуемые отличаются от послушных по какому-либо важному признаку (возможно, у них еще до начала эксперимента хуже с сердцем?), мы не можем предсказать, как повлиял бы на них препарат, если бы они следовали инструкциям.

В этой ситуации у нас получается каузальная диаграмма, как на рис. 51. Переменная *назначение* (*Z*) будет равна 1, если в результате случайного жребия пациенту выпало принимать препарат, и равна 0, если он будет получать плацебо. Переменная *прием* будет равна 1, если пациент принимал препарат, и равна 0, если нет. Для удобства мы также используем бинарное определение для переменной *холестерин* с исходом 1, если холестерин пациента снизился на определенную фиксированную величину. Обратите внимание, что в этом случае наши переменные бинарные, а не численные. Это сразу же означает, что мы не можем в этом случае использовать линейную модель, а следовательно, и применять формулу для инструментальных переменных, выведенную ранее. Тем не менее в таких случаях мы часто заменяем допущение о линейности более слабым допущением о монотонности, которое я объясню ниже.

Но прежде чем мы сделаем это, давайте удостоверимся, что другие необходимые для инструментальных переменных допущения в данном случае действительны. Во-первых, является ли инструментальная переменная *Z* независимой от конфаундера? Рандомизация *Z* гарантирует ответ «да» (как мы видели в главе 4, рандомизация — замечательный способ удостовериться, что на переменную не влияют никакие конфаундеры. Есть ли прямой путь от *Z* к *Y*? Здравый смысл говорит, что получение конкретного случайного номера (*Z*) никак не воздействует на уровень холестерина (*Y*), поэтому ответ «нет». Наконец, есть ли сильная ассоциация между *Z* и *X*? На этот раз нужно смотреть на сами данные, и ответ снова «да». Эти три вопроса мы должны задавать всегда перед применением инструментальных переменных.

В нашем примере ответы очевидны, но мы не должны закрывать глаза на тот факт, что, отвечая на них, мы используем каузальную интуицию, которая запечатлевается, сохраняется и проясняется в виде каузальной диаграммы. Табл. 10 пока-

зывает наблюдаемую частоту исходов X и Y. Так, для 91,9% людей, которым не назначался препарат, наблюдается исход $X = 0$ (не принимали препарат) и $Y = 0$ (уровень холестерина не упал). В этом есть смысл. У оставшихся 8,1% исход был $X = 0$ (не принимали препарат) и $Y = 1$ (уровень холестерина упал). Очевидно, их состояние улучшилось по каким-то иным причинам. Обратите внимание также, что в таблице два нуля: не было никого, кому назначили бы плацебо ($Z = 0$), но кто тем не менее каким-нибудь образом добыл бы препарат ($X = 1$). В хорошо организованном рандомизированном исследовании, особенно в области медицины, где только у врачей есть доступ к экспериментальному препарату, дело обычно обстоит именно так. Предположение, что в выборке нет индивидов, у которых $Z = 0$, а $X = 1$, называется монотонностью.

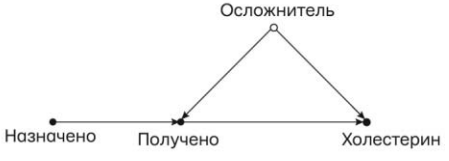


Рис. 51 Каузальная диаграмма для рандомизированного контролируемого исследования с неподчинением

Теперь посмотрим, как оценить влияние приема препарата. Сначала рассмотрим наихудший сценарий: никому из непослушных испытуемых не стало бы лучше, даже если бы они принимали препарат. В этом случае все люди, которые теоретически, принимая препарат, могли бы улучшить свое состояние, уже сосредоточены в той группе в 47,3%, в которой испытуемые реально принимали его и реально улучшили свое здоровье.

Но нам нужно скорректировать эту оценку по эффекту плацебо, данные по которому в третьем ряду таблицы. Из людей, которым назначили плацебо и которые его принимали, показатели улучшились у 8,1%. Таким образом, чистые показатели, выходящие за уровень плацебо, составляют 47,3—8,1 = 39,2%.

Теперь рассмотрим наилучший сценарий, при котором все люди, не принимавшие назначенный им препарат, снизили бы холестерин, если бы послушались. В этом случае мы прибавляем к 31,5% непослушных 7,3 и к этому только что подсчитанный нижний порог в 39,2, получая сумму в 78,0%.

Таким образом, даже при наихудшем сценарии, в котором конфаундеры действуют полностью противоположно эффекту препарата, мы все же вправе сказать, что этот препарат улучшает уровень холестерина для 39% популяции. В наилучшем сценарии, когда конфаундер действует «на руку» препарату, улучшение будет наблюдаться для 78% популяции.

Таблица 11. Данные эксперимента с холестирамином

Исход	Не назначали препарат ($Z = 0$)	Назначали препарат ($Z = 1$)
$X = 0, Y = 0$	0,919	0,315
$X = 1, Y = 0$	0	0,139
$X = 0, Y = 1$	0,081	0,073
$X = 1, Y = 1$	0	0,473

Даже несмотря на то, что границы довольно далеко отстоят друг от друга из-за большого числа испытуемых, не подчинившихся условиям эксперимента, исследователи могут категорически утверждать, что препарат эффективно достигает своей цели.

Эта стратегия рассмотрения наихудшего и наилучшего сценариев обычно дает нам некоторый диапазон оценок. Очевидно, что желательно было бы получить точечную оценку, как в случае линейных зависимостей. Существуют способы при необходимости сузить этот диапазон, а в некоторых случаях даже добиться точечных оценок. Так, если вас интересует только «послушная» часть популяции (те, кто будет принимать X тогда и только тогда, когда его им назначат), вы можете вывести точечную оценку, известную как локальный средний эффект лечения (LATE). В любом случае я надеюсь, этот пример пока-

жет, что наши руки ничто не связывает, даже если мы покидаем мир линейных моделей.

Методы инструментальных переменных продолжали развиваться с 1984 года, и одна конкретная версия стала очень популярной: менделева рандомизация. Вот вам доказательство. Хотя влияние ЛПНП, или «плохого» холестерина, сегодня хорошо известно, далеко не все однозначно понятно в случае «хорошего» холестерина — липопротеина высокой плотности, ЛПВП. Ранние исследования на основе наблюдений, скажем Фрамингемские исследования сердца в конце 70-х годов XX века, предположили, что ЛПВП обладают защитными свойствами, предохраняя от инфаркта. Однако ЛПВП обычно встречается вместе с ЛПНП, так как же нам узнать, какой из липидов на самом деле является каузальным агентом?

Чтобы ответить на этот вопрос, предположим, что нам известен ген, на уровень ЛПНП не влияющий, но благодаря которому у людей выше уровень ЛПВП. Тогда нам удастся нарисовать каузальную диаграмму, как на рис. 52, где я изобразил переменную *образ жизни* как потенциальный конфаундер. Вспомним, что всегда лучше, как в примере доктора Сноу, использовать инструментальную переменную, которая рандомизирована. В таком случае к ней не идут каузальные стрелки. По этой причине ген — отличная инструментальная переменная. Наши гены рандомизированы в момент зачатия так, словно Грегор Мендель дотянулся рукой с небес и случайно назначил одним людям ген высокого риска, а другим — ген низкого риска инфаркта. Отсюда возник термин «менделева рандомизация».

Может ли здесь быть стрелка, идущая в обратном направлении, от гена ЛПВП к образу жизни? Здесь нам снова требуется «расследование на подметках ботинок» и каузальное мышление. Ген ЛПВП мог бы влиять на образ жизни людей только в том случае, если бы они изначально знали, какая версия гена им досталась — с высоким уровнем ЛПВП или с низким. Но до 2008 года такие гены были неизвестны, да и сегодня у людей обычно нет доступа к подобной информации о себе. Поэтому весьма вероятно, что такой стрелки не существует.

По крайней мере два исследования холестеринового вопроса использовали этот подход менделевой рандомизации. В 2012 году масштабное совместное исследование, возглавляемое Секаром Катиресаном из Массачусетской больницы общего профиля, показало, что никаких преимуществ от более высокого уровня ЛПВП не наблюдается. Правда, эти исследователи обнаружили, что ЛПНП очень значительно влияет на риск инфаркта. Согласно их результатам, сокращение уровня ЛПНП на 34 мг/дл сокращает ваши шансы получить инфаркт на 50%. Поэтому снижение «плохого» холестерина, будь то при помощи диеты, физических упражнений или статинов, — это хорошая идея. Тем не менее повышение уровня «хорошего» холестерина, что бы вам там ни говорили производители рыбьего жира, похоже, никак не влияет на риск получить инфаркт.

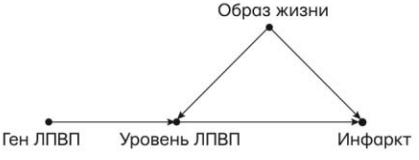


Рис. 52. Каузальная диаграмма для примера с менделевой рандомизацией

Как обычно, к вышесказанному есть и оговорка. Второе исследование, опубликованное в том же году, отметило, что у людей с менее опасным вариантом гена ЛПНП низкий уровень холестерина сохраняется в течение всей их жизни. Менделева рандомизация говорит нам, что, сокращая ваш уровень ЛПНП на 34% в течение всей вашей жизни, снижает ваш риск умереть от инфаркта наполовину. Но статины не способны снижать ваш уровень «плохого» холестерина подобным образом — они действуют только с того дня, с которого вы начали их принимать. Если вам 60 лет, у ваших артерий уже 60 лет износа. По этой причине весьма вероятно, что менделева рандомизация переоценивает истинную пользу статинов.

Однако, начав снижать уровень холестерина смолоду — посредством диеты, или физкультуры, или даже статинов, — вы спустя время добьетесь значительных результатов.

С точки зрения каузального анализа вышесказанное преподает нам хороший урок: в любом исследовании интервенций нам следует выяснить, действительно ли та переменная, которой мы реально манипулируем (например, уровень ЛПНП в течение жизни), — это та же самая переменная, про которую мы думаем, что манипулируем ей (уровень ЛПНП в настоящий момент). Это входит в «искусный допрос природы».

Подытожим: инструментальные переменные — важный инструмент, они помогают нам выявить каузальную информацию, выходящую за пределы *do*-исчисления. Последнее настаивает на точечных оценках, а не на неравенствах и не работает в случаях вроде приведенного на рис. 52, в котором все, что нам удается получить, — это неравенства. Кроме того, важно понимать, что *do*-исчисление намного более гибко, чем метод инструментальных переменных. В *do*-исчислении нам не нужно делать никаких допущений относительно природы функций в каузальных моделях. Но если мы способны научно обосновать допущение о монотонности или линейности такой функции, тогда такой более специализированный инструмент, как инструментальные переменные, стоит принять к рассмотрению.

Методы инструментальных переменных можно распространить за пределы простых моделей из четырех переменных, как на рис. 49, но без опоры на каузальные диаграммы не получится уйти далеко. Например, в некоторых случаях несовершенная инструментальная переменная (т.е. такая, которая не вполне независима от конфаундера) используется после введения поправок по разумно подобранному набору вспомогательных переменных, блокирующих пути между инструментальной переменной и конфаундером. Мой бывший студент Карлос Брито, ныне профессор в Федеральном университете Сеары в Бразилии, полностью развил эту идею превращения неинструментальных переменных в инструментальные.

Вдобавок к этому Брито изучил множество случаев, в которых целый набор переменных успешно используется в качестве

инструментальной. Хотя идентификация инструментальных наборов выходит за пределы *do*-исчисления, при этом все же используются каузальные диаграммы. Для исследователей, понимающих такой язык, возможные схемы экспериментов весьма разнообразны: им не нужно ограничивать себя только четырехпеременными моделями, показанными на рис. 49, 51 и 52. Наши возможности ограничены только нашим воображением.

Пути было два, и мир был широк
Однако я раздвоиться не мог.[2]

Знаменитые строчки Роберта Фроста отражают глубокое понимание поэтом контрфактивного. Мы не можем странствовать по двум дорогам одновременно, однако наш разум наделен способностями судить, что произошло бы, если бы мы выбрали другой путь. Вооружившись этим суждением, к концу поэмы Фрост оказывается доволен своим выбором, понимая, что «все остальное не играет роли».

---

[2] Ориг. «The road not taken» — Robert Frost, перевод Г. Кружкова «Неизбранная дорога» (источник: stihi.rU)

Глава 8

Контрфактивные суждения: глубинный анализ миров, которые могли бы существовать

Если бы нос Клеопатры был немного короче,
то изменился бы лик всей Земли.
Блез Паскаль

Готовясь перейти на следующий уровень Лестницы Причинности, давайте обобщим, что мы узнали на втором уровне. Мы видели, что существуют несколько способов гарантировать эффект интервенции в разных контекстах и при разных условиях. В главе 4 мы обсудили рандомизированные контролируемые исследования, широко цитируемый золотой стандарт медицинских испытаний. Также мы рассмотрели методы, подходящие для наблюдательных исследований, в который испытуемая и контрольная группы выбирается непроизвольно. Если нам удается измерить все переменные, которые блокируют черные входы, формула поправки черного входа используется, чтобы получить необходимый эффект. Если мы найдем путь через парадный ход, закрытый от конфаундеров, то сможем использовать поправку парадного хода. Если же мы готовы принять линейность или монотонность, то применим инструментальные переменные (предполагая, что соответствующая переменная найдется на диаграмме или будет создана экспериментально). А действительно предприимчивые исследователи проложат

другие маршруты к вершине горы Интервенции, используя *do*-исчисление или его алгоритмическую версию.

Во всех этих случаях мы имели дело с эффектом воздействия на исследуемую выборку или на типичного индивида, взятого из этой выборки (усредненный эффект от причинно-следственной взаимосвязи). Но пока мы упустили из обсуждения причинно-следственную связь на личном уровне — уровне отдельных событий или индивидов. Одно дело — сказать, что курение вызывает рак, но совсем другое — заявить, что ваш дядя Джо, который выкуривал по пачке сигарет 30 лет подряд, остался бы в живых, если бы не курил. Разница одновременно очевидна и глубока: никого из тех, кто, подобно дяде Джо, курил 30 лет и умер, нельзя наблюдать в альтернативной реальности, где они не курили 30 лет.

Ответственность и вина, сожаление и доверие — эти понятия служат ходовой валютой в причинно-следственных рассуждениях. Чтобы как-то их истолковать, у нас должна быть возможность сравнить то, что действительно случилось, с тем, что случилось бы гипотетически в какой-то альтернативной ситуации. Как я утверждал в главе 1, способность представлять альтернативные, несуществующие миры отделила нас от протолюдей и, более того, от всех остальных существ на планете. Любое другое существо видит то, что есть. Наш дар, который порой может быть проклятием, — видеть то, что могло бы быть.

Эта глава показывает, как использовать данные наблюдений и экспериментов, чтобы добывать информацию о контрфактивных сценариях. Она объясняет, как представлять причины индивидуального уровня на диаграммах причинности — задача, которая вынудит нас объяснить некоторые составные элементы диаграмм, о которых мы еще не говорили. Также я коснусь тесно связанного с ними понятия возможных результатов, или модели Неймана — Рубина, изначально предложенной в 1920-х годах Ежи Нейманом, польским статистиком, который позже стал профессором Калифорнийского университета в Беркли. Однако этот подход к причинному анализу получил развитие только в середине 1970-х, когда Дональд Рубин начал писать о потенциальных результатах.

Я покажу, как контрфактивность возникает естественным образом в контексте, описанном в последних нескольких главах, — в путевых диаграммах Сьюалла Райта и их расширении на структурные модели причинности (*Structural Causal Models; SCM*). Мы получили хорошее представление об этом в главе 1 на примере расстрельной команды, в котором было показано, как отвечать на контрфактивные вопросы вроде: «Остался бы заключенный в живых, если бы стрелок А не стрелял?». Я сравню, как контрфактивные суждения определяются в парадигме Неймана — Рубина и в SCM, где они пользуются преимуществом каузальных диаграмм. Рубин на протяжении многих лет уверенно утверждал, что диаграммы бесполезны. Итак, мы посмотрим, как изучающие его модель вынуждены ориентироваться в причинных проблемах с завязанными глазами, не имея способа, чтобы представить причинное знание или вывести его проверяемые следствия.

Наконец, мы рассмотрим два вида ситуаций, в которых использование контрфактивных суждений абсолютно необходимо. Десятилетиями или даже столетиями юристы использовали относительно простой метод, чтобы проверить виновность подсудимого — принцип sine q*U*a non («то, без чего невозможно»): вред не был бы нанесен, если бы подсудимый не совершил действие. Мы увидим, как контрфактивный язык распознает это неуловимое понятие и позволяет оценить вероятность того, что обвиняемый виновен.

Затем я расскажу, как контрфактивные суждения могут быть применены к климатическим изменениям. До недавнего времени климатологам было трудно и неловко отвечать на вопросы вроде: «Вызвало ли глобальное потепление этот шторм (или эту жару, или эту засуху)?». Традиционный ответ был таким: отдельные погодные явления нельзя приписывать глобальному изменению климата. Однако этот ответ кажется довольно уклончивым и даже способствует укреплению безразличия общественности к названной проблеме.

Контрфактивный анализ позволяет климатологам делать гораздо более четкие и точные утверждения, чем раньше. Однако они требует внести небольшие правки в нашу повседнев-

ную лексику. Будет полезно различать три типа причинности: *необходимую*, *достаточную* и *необходимую и достаточную* (необходимая причинность и есть «то, без чего невозможно»). Используя эти выражения, ученый-климатолог скажет: «Существует 90%-ная вероятность того, что антропогенные изменения климата были необходимой причиной этой жары» или «Существует 80%-ная вероятность того, что изменений климата достаточно, чтобы вызвать такую сильную жару, по крайней мере один раз в 50 лет». Первое предложение связано с атрибуцией: кто виноват в необычной жаре? Второе связано с программой действий. Оно сообщает, что лучше подготовиться к таким периодам жары, потому что рано или поздно они придут. Оба эти утверждения более информативны, чем пожимания плечами и отказ что-то говорить о причинах отдельных погодных явлений.

От Фукидида и Авраама до Юма и Льюиса

Учитывая тот факт, что контрфактивные рассуждения — часть ментального аппарата, которая делает нас людьми, неудивительно, что мы обнаруживаем их существование в настолько давние времена, насколько захотим углубиться. Так, в «Истории Пелопонесской войны» Фукидида, древнегреческого историка, которого часто называют пионером научного подхода к истории, описывает цунами, которое произошло в 426 году до н.э.: «Примерно в то время, когда подобные землетрясения были столь часты, море при Оробиях, что на Эвбее, отступило от тогдашней береговой линии, вернулось в виде огромной волны, поглотило значительную часть города, а потом ушло, но не до конца, и там, где раньше была суша, теперь море. Все же обитатели города, которые не успели подняться на высокие места, погибли… По моему мнению, причину этого явления необходимо искать в землетрясении. Там, где оно сильнее всего, море отходит от берега, а потом внезапно возвращается

с удвоенной силой, вызывая наводнение. Не могу представить, как такое могло бы произойти без землетрясения».

Это, несомненно, примечательный отрывок, если учесть эпоху, когда он был написан. Во-первых, точность наблюдений Фукидида сделала бы честь любому современному ученому, тем более что он работал в эпоху, когда не было ни спутников, ни видеокамер, ни круглосуточных новостных служб, передающих картины катастрофы в реальном времени. Во-вторых, в его историческое время стихийные бедствия регулярно приписывали воле богов. Его предшественник Гомер или современник Геродот непременно приписали бы это событие гневу Посейдона либо какого-то иного божества. Но Фукидид предлагает причинную модель без сверхъестественных процессов: землетрясение оттягивает море, но потом оно возвращается и затапливает землю. Последнее предложение в этой цитате особенно интересно, потому что оно объясняет необходимую причинность: если бы не землетрясение, цунами бы не произошло. Эта контрфактивная оценка переводит землетрясение из разряда чего-то, предшествующего цунами, в его действительную причину.

Еще один завораживающий и показательный пример контрфактивных рассуждений встречается в Книге Бытия в Библии. Авраам спрашивает у Бога о намерении последнего уничтожить города Содом и Гоморру в наказание за нечестивое поведение их жителей.

«И подошел Авраам, и сказал: неужели Ты погубишь праведного с нечестивым?

Может быть, есть в этом городе пятьдесят праведников? Неужели Ты погубишь, и не пощадишь места сего ради пятидесяти праведников в нем? <…>

Господь сказал: если Я найду в городе Содоме пятьдесят праведников, то Я ради них пощажу все место сие».

Но на этом история не заканчивается. Авраам не удовлетворен и спрашивает Господа: а что, если есть лишь 45 праведников? Или 40? Или 30? Или 20? Или даже 10? Каждый

раз он получает утвердительный ответ, и Бог в конце концов заверяет его, что сохранит Содом даже ради 10 праведников, если получится столько найти.

Чего пытается добиться Авраам, торгуясь и выпрашивая? Конечно же, он не сомневается в способностях Бога считать. И конечно, Авраам знает, сколько именно праведников живет в Содоме. В конце концов, он вездесущий.

Зная о покорности, преданности Авраама, трудно поверить, что с помощью этих вопросов он хотел уговорить Бога поменять решение. Напротив, они нужны самому Аврааму, чтобы разобраться. Он рассуждает так же, как это сделал бы современный ученый, который пытается понять законы, управляющие коллективным наказанием. За какой уровень нечестивости полагается уничтожение? Хватит ли 30 праведников, чтобы спасти город? А 20? У нас не будет настоящей модели причинности без такой информации. Современный ученый мог бы назвать это кривой «доза — эффект» или пороговым эффектом.

В то время как Фукидид и Авраам подступались к контрфактивным суждениям на базе отдельных случаев, греческий философ Аристотель исследовал более общие аспекты причинности. В своем типично систематическом стиле Аристотель разработал целую классификацию первопричин, в которую вошли форма, материя, цель и перводвижитель. Например, бронза и ее свойства служат причиной для очертаний статуи, из которой та отлита; такую же статую нельзя было бы сделать из пластилина. Однако Аристотель нигде не рассуждает о причинности, используя контрфактивные суждения, поэтому его изобретательной классификации не хватает простой ясности, которую мы видим в отчете Фукидида о причине цунами.

Чтобы найти философа, который поместил контрфактивность в самое сердце причинности, нужно переместиться во времени к Дэвиду Юму, шотландскому философу и современнику Томаса Байеса. Юм отрицал классификацию Аристотеля и настаивал на единственном определении причинности. Однако он обнаружил, что эта дефиниция ускользает от него, и, более того, понял, что не способен выбрать одно из двух разных определений. Позже они превратятся в два несовмести-

мых течения, и по иронии судьбы представители обоих будут говорить, что источником для них был Юм!

В «Трактате о человеческой природе» Юм отрицает, что любые два объекта имеют внутренние свойства, или «силы», которые делают одного причиной, а другого — следствием. По его мнению, причинно-следственное отношение — это исключительно продукт нашей памяти и опыта. «Таким образом, мы помним, что наблюдали разновидность объекта, которую называем *огнем*, и испытывали разновидность ощущения, которое называем *жаром*, — пишет он. — Подобным образом мы вызываем в уме их постоянную связь во всех примерах из прошлого. Без дальнейших церемоний мы называем одно *причиной*, а другое — *следствием* и выводим существование одного из существования другого (курсив — Д. Юма)». Сегодня это известно как объяснение причинности через «регулярность».

Этот отрывок поражает безапелляционностью. Юм отбрасывает второй и третий уровни Лестницы Причинности и утверждает, что первый уровень, наблюдение, — это все, что нам нужно. Как только огонь и жара попадутся нам вместе достаточно много раз (учтите, что огонь должен предшествовать во времени), мы согласимся назвать огонь причиной жары. Как и большинство статистиков XIX века, Юм в 1739 году, кажется, счастлив считать причинность всего лишь разновидностью корреляции.

К чести Юма надо сказать, что он не был удовлетворен этим определением. Девять лет спустя в «Исследовании о человеческом познании» он написал нечто совершенно иное: «Мы можем определить причину *как объект, за которым следует другой объект, когда за всеми объектами, схожими с первым, следуют объекты, схожие со вторым.* Или, другими словами, *если бы не было первого объекта, второй никогда не существовал бы».* Первое предложение, версия, где А постоянно наблюдают вместе с В, просто повторяет определение через регулярность. Но к 1748 году, кажется, у него появились определенные опасения и он решил кое-что исправить. Будучи настоящими историками-вигами, мы можем понять почему. Согласно его более раннему определению, кукареканье петуха

должно вызывать рассвет. Чтобы справиться с этой трудностью, он добавляет второе определение, на которое даже не намекал в более ранней книге — контрфактивное определение: «если бы не было первого объекта, второй никогда не существовал бы».

Обратите внимание на то, что второе определение в точности повторяет то, которое использовал Фукидид, обсуждая цунами при Оробиях. Контрфактивное определение также объясняет, почему мы не считаем кукареканье петуха причиной рассвета. Мы знаем, что, если в какой-то день петух заболеет или капризно откажется кукарекать, солнце встанет все равно.

Хотя Юм пытается представить два эти определения как одно, вставляя невинное «другими словами», второй вариант полностью отличается от первого. Он очевидно подразумевает контрфактивность, а значит, находится на третьем уровне Лестницы Причинности. В то время как регулярные вещи можно наблюдать, контрфактивные можно только вообразить.

Стоит задуматься на минуту о том, почему Юм решил определять причины через контрфактивные суждения, а не наоборот. Определения нужны, чтобы свести более сложное понятие к более простому. Юм предполагает, что его читатели воспримут утверждение «если бы не было первого объекта, второй никогда не существовал бы» как менее двусмысленное, чем «первый объект стал причиной второго». Он абсолютно прав. Это последнее утверждение способно вызвать самые разные бесплодные метафизические измышления о том, какие качества или силы, присутствующие в первом объекте, вызывает второй. Первое утверждение заставляет нас пройти простой тест в уме: вообразите мир без землетрясения и спросите, было ли в нем цунами. Мы делали подобные умозаключение с детства, а люди как биологический вид начали еще во времена Фукидида (и, вероятно, задолго до этого).

Тем не менее философы игнорировали второе определение Юма большую часть *XIX* и *XX* веков. Контрфактивные суждения, все эти «было бы», всегда казались слишком скользкими и неопределенными, чтобы удовлетворить ученых. Вместо этого философы пытались спасти первое определение Юма

с помощью теории вероятностной причинности, которую мы обсуждали в главе 1.

Философ Дэвид Льюис бросил вызов традиционным представлением. В книге 1973 года «Возможные миры» (*Counterfactuals*) он призвал вообще отказаться от обращения к регулярности и интерпретировать утверждение «*A* вызвало *B*» как «*B* не произошло бы, если бы не *A*». Льюис спрашивал: «Почему бы не принимать контрфактивные суждения буквально — как утверждения о возможных альтернативах реальной ситуации?»

Как и Юм, Льюис был явно впечатлен тем фактом, что люди делают контрфактивные суждения без особой подготовки — быстро, постоянно и не напрягаясь. Мы способны приписывать им значения истинности и вероятности так же уверенно, как делаем это для фактических утверждений. По мнению Льюиса, мы совершаем это, представляя «возможные миры», в которых контрфактивные утверждения верны.

Когда мы говорим: «У Джо прошла бы головная боль, если бы он принял аспирин», по Льюису, мы утверждаем, что существуют другие возможные миры, в которых Джо таки принял аспирин и его головная боль прошла. Льюис утверждал, что мы оцениваем контрфактивные суждения, сравнивая мир, в котором он не выпил лекарство, с самым похожим миром, в котором Джо все-таки его выпил. Не обнаружив в этом мире головной боли, мы объявляем, что контрфактивное суждение верно. «Наиболее похожий» — это ключевой момент. В других «возможных мирах» его головная боль могла и не пройти, например в мире, где он принял аспирин, а потом ударился головой о дверь в ванной. Но в выбранном нами мире сложилась благоприятная ситуация. Среди всех возможных миров, в которых Джо принял аспирин, ближе всех нашему оказался бы не тот, где он ударился головой, а тот, где его боль прошла.

Многие критики Льюиса ухватились за экстравагантность его утверждений о буквальном существовании множества других миров. «Мистера Льюиса когда-то окрестили „одержимым модальным реалистом" за идею о том, что любой логически возможный мир, о котором можно помыслить, действительно существует, — говорится в его некрологе, выпущенном

в «Нью-Йорк таймс» в 2001 году. — Он считал, например, что существует мир с говорящими ослами».

Но я думаю, что эти критики (и, возможно, сам Льюис) упустили самый важный момент. Нет нужды спорить о том, существуют ли такие миры как физические или даже метафизические сущности. Если мы хотим объяснить, какой смысл люди вкладывают во фразу «А вызывает В», достаточно постулировать, что они способны генерировать в голове альтернативные миры, оценивая, какой мир «ближе» к нашему, и, что самое важное, делать это последовательно, с целью прийти к консенсусу.

Конечно, мы не могли бы обсуждать контрфактивные ситуации, если бы «ближе» для одного человека было бы «дальше» для другого. С этой точки зрения, призывая «воспринимать возможные миры как таковые», Льюис предлагал не ударяться в метафизику, а проявить внимание к удивительному единообразию в архитектуре человеческого разума.

Как лицензированный философ-виг, я способен довольно хорошо объяснить эту особенность: она проистекает из того, что мы переживаем один и тот же мир и разделяем одну и ту же ментальную модель его причинной структуры. Мы говорили об этом еще в главе 1. Общие ментальные модели связывают нас в сообщества. Следовательно, мы судим о близости не по метафизическому понятию сходства, а по тому, насколько мы должны разобрать и изменить общую модель, прежде чем она удовлетворит заданному гипотетическому условию, противоречащему действительности (Джо не принял аспирин).

В структурных моделях мы делаем очень похожие вещи, хотя больше украшаем их математическими деталями. Мы оцениваем выражения вроде «если бы X был X» так же, как обрабатываем интервенции $do\ (X = x)$, удаляя стрелки в диаграмме причинности или уравнение в структурной модели. Это описывается как минимальное изменение причинно-следственной диаграммы, необходимое, чтобы гарантировать равенство X и x. В этом отношении структурные контрфактивные модели совместимы с идеей Льюиса о максимально похожем возможном мире.

Структурные модели также предлагают решение загадки, о которой Льюис умалчивал: каким образом люди представляют возможные миры и вычисляют ближайший, когда число этих возможностей слишком велико для человеческого мозга? Специалисты по компьютерным наукам называют это проблемой представления. Предполагается, что у нас есть некий крайне экономичный код, чтобы иметь дело с таким количеством миров. Могут ли структурные модели в той или иной форме быть этим коротким путем? Думаю, это весьма вероятно по двум причинам. Во-первых, структурные причинно-следственные модели работают, и у них просто нет конкурентов с такими же чудесными свойствами. Во-вторых, они были созданы на основе байесовских сетей, которые, в свою очередь, были смоделированы на базе сделанного Дэвидом Румельхартом описания, как сообщения передаются в мозгу. Нетрудно предположить, что 40 тысяч лет назад люди адаптировали механизмы, которые уже существовали в мозге для распознавания образов, чтобы использовать их для причинно-следственных рассуждений.

Философы, как правило, оставляют психологам право делать утверждения о том, как работает человеческий разум. Это объясняет, почему вышеперечисленные вопросы не рассматривались до недавнего времени. Однако исследователи искусственного интеллекта не могли ждать. Они хотели создать роботов, которые были бы способны общаться с людьми об альтернативных сценариях, доверии и вине, ответственности и сожалении. Это все контрфактивные представления, для которых исследователи ИИ должны были найти механизмы, чтобы у них появился минимальный шанс создать так называемый сильный ИИ — интеллект, подобный человеческому.

Такова была моя мотивация, когда я начал изучать контрфактивный анализ в 1994 году (вместе с моим студентом Алексом Балке). Неудивительно, что алгоритмизация контрфактивных суждений произвела больший фурор в мире ИИ и когнитивной науке, чем в философии. Философы склонны рассматривать структурные модели как один из вариантов применения логики возможных миров, представленной Льюисом. Осмелюсь предположить, что их роль гораздо более велика. Логическая

пустота представления — это метафизика. Диаграммы причинно-следственных связей с простыми правилами следования стрелкам или удаления стрелок должны быть близки к тому, как наш мозг представляет контрфактивные суждения.

Этому утверждению пока суждено оставаться недоказанным, но в результате долгого процесса контрфактивные суждения утратили мистический флер. Мы понимаем, как люди справляются с ними и готовы вооружить роботов возможностями, аналогичными тем, которые наши предки приобрели 40 тысяч лет назад.

Потенциальные результаты, структурные уравнения и алгоритмизация контрфактивных утверждений

Всего через год после выхода книги Льюиса и независимо от него Дональд Рубин начал работать над серией статей, в которой представлены потенциальные результаты как язык для постановки вопросов о причинности. Рубин, в то время статистик службы тестирования в образовании, в одиночку нарушил молчание о причинно-следственных связях, которое сохранялось в статистике на протяжении 75 лет, и легитимировал концепцию контрфактивности в глазах многих ученых-медиков. Важность этого достижения нельзя переоценить. Исследователи получили гибкий язык для выражения почти любых каузальных вопросов, которые они могли бы задать и о группе людей, и об отдельных индивидах.

В каузальной модели Рубина потенциальный результат для переменной Y — это просто «значение, которое Y принял бы для отдельного u, если бы X было присвоено значение x». Здесь очень много слов, поэтому часто удобнее записать эту величину более компактно: $Y_X = {}_X(u)$. Нередко мы сокращаем это до $Y_X(u)$, если из контекста очевидно, какой переменной присваивается значение X.

Чтобы оценить смелость такого рода записи, стоит отвлечься от символов и подумать о том, что за ними стоит. Записывая символ Y_x, Рубин утверждал, что Y определенно приобрел бы какое-то значение, если бы X был равен x, и этот факт объективно реален в той же степени, что и значение, которое Y получил на самом деле. Если вы не согласны с этим допущением (а я уверен, что Гейзенберг с ним не согласился бы), то не сможете использовать потенциальные результаты. Также обратите внимание, что потенциальный, или контрфактивный, результат определяется на уровне отдельного человека, а не группы.

Впервые потенциальные результаты появились в магистерской диссертации Ежи Неймана, написанной в 1923 году. Нейман, потомок польских аристократов, вырос в изгнании в России, где получил очень сильное математическое образование, и оказался на родине только в 1921 году, когда ему было 27 лет. Он хотел продолжить исследования в области чистой математики, но ему было легче найти работу статистика. Как и Р. Э. Фишер в Англии, он провел первое статистическое исследование в сельскохозяйственном институте и оказался слишком высококвалифицированным для этой работы. Он был не только единственным статистиком в институте, но и единственным человеком в стране, который думал о статистике как о научной дисциплине.

Первое упоминание о потенциальных результатах Нейман сделал в контексте сельскохозяйственного эксперимента, где нижний индекс представляет «неизвестный потенциальный урожай i-й разновидности [данного семени] на соответствующем участке». Тезис оставался неизвестным и не переводился на английский до 1990 года. Однако сам Нейман неизвестным не остался. Он договорился провести год в статистической лаборатории Карла Пирсона в Университетском колледже Лондона, где подружился с сыном Пирсона Эгоном. Эти двое поддерживали связь в течение следующих семи лет, и их сотрудничество принесло большие плоды: подход Неймана — Пирсона к статистической проверке гипотез стал важной вехой, о которой узнает каждый начинающий студент-статистик.

В 1933 году длительное автократическое правление Карла Пирсона наконец подошло к концу с его уходом на пенсию, и Эгон стал его естественным преемником или оказался бы таковым, если бы не единственная проблема в виде Р. Э. Фишера, к тому времени самого известного статистика в Англии. Университет предложил уникальное и катастрофическое решение, разделив территорию Пирсона на кафедру статистики (Эгон Пирсон) и кафедру евгеники (Фишер). Эгон, не теряя времени, нанял своего польского друга. Нейман прибыл в 1934 году и почти сразу же схлестнулся с Фишером.

Фишер уже рвался в бой. Он знал, что является ведущим статистиком мира и во многом практически изобрел этот предмет, однако ему было запрещено преподавать на отделении статистики. Отношения были необычайно напряженными. «Комнату преподавателей тщательно делили, — пишет Констанс Рид в своей биографии Неймана. — Группа Пирсона пила чай в 4 часа; в 4:30, когда они благополучно удалялись, десантировалась группа Фишера».

В 1935 году Нейман прочитал в Королевском статистическом обществе лекцию под названием «Статистические проблемы сельскохозяйственных экспериментов», в которой подверг сомнению некоторые методы Фишера, а также между прочим обсудил идею потенциальных результатов. Когда Нейман закончил, Фишер встал и заявил, что «надеялся, что статья доктора Неймана будет посвящена теме, с которой автор полностью знаком».

«[Нейман] утверждал, что Фишер был неправ, — писал Оскар Кемпторн много лет спустя об этом инциденте. — Это было непростительное преступление — Фишер никогда не ошибался, и предположение о том, что это, возможно, расценивалось как вооруженное нападение. Всякий, кто не принимал писания Фишера как данную Богом истину, был в лучшем случае глупцом, а в худшем — злодеем». Несколько дней спустя Нейман и Пирсон увидели всю силу его гнева, когда вечером пришли на факультет и обнаружили разбросанные по полу деревянные модели Неймана, которыми он иллюстрировал свою лекцию.

Они пришли к выводу, что только Фишер мог устроить эти разрушения.

Хотя сейчас этот приступ ярости покажется забавным, позиция Фишера имела серьезные последствия. Конечно, он не был способен обуздать свою гордость и использовать запись потенциального результата, предложенную Нейманом, хотя это помогло бы ему позже с проблемами медиации. Отсутствие языка потенциальных результатов привело его и многих других к так называемой ошибке посредничества, которую мы обсудим в главе 9.

На этом этапе некоторые читатели, вероятно, все еще считают концепцию контрфактивности несколько мистической, поэтому я хотел бы показать, как некоторые последователи Рубина делают выводы о потенциальных результатах, и противопоставить этот безмодельный подход структурной причинно-следственной модели.

Представим, что мы изучаем конкретную компанию, пытаясь понять, что сильнее влияет на зарплату сотрудника — образование или многолетний стаж. Мы собрали данные о существующих зарплатах в этой компании и записали их в табл. 12. Условимся, что *EX* — стаж, *ED* — образование, *S* — зарплата. Также для простоты предположим, что существуют три уровня: 0 = средняя школа, 1 = высшее образование, 2 = ученая степень. Таким образом, $S_{ED} = 0(u)$ или $S_0(u)$ представляет собой зарплату человека u, если u окончил среднюю школу, но не университет, а $S_1(u)$ представляет зарплату u, если бы тот окончил университет. Типичный контрфактивный вопрос, который можно было бы задать, звучит так: какой была бы зарплата Элис, если бы у нее было высшее образование? Другими словами, чему равна S_1 (Элис)?

Первое, на что следует обратить внимание в табл. 12, — это отсутствующие данные, отмеченные вопросительными знаками. Для одного и того же человека нельзя увидеть более одного потенциального результата. Несмотря на всю очевидность, это важное утверждение. Статистик Пол Холланд однажды назвал его фундаментальной проблемой причинного вывода, и название прижилось. Если бы мы могли заполнить клетки

с вопросительными знаками, то ответили бы на все наши вопросы о причинности.

Я никогда не был согласен с представлением Холланда об отсутствующих данных в табл. 12 как о «фундаментальной проблеме», возможно, потому, что я редко представлял проблемы причинности в виде таблицы. Но если подойти к делу фундаментально, становится понятно, что его подход чреват огромными заблуждениями, что мы вскоре увидим. Обратите внимание, что, помимо декоративных заголовков последних трех столбцов, табл. 12 полностью лишена каузальной информации о *ED*, *EX* и *S*, например о том, влияет образование на заработную плату или наоборот. Хуже того, она не позволяет нам представлять такую информацию, даже когда она доступна. Но статистикам, которые видят фундаментальную проблему в отсутствии данных, такая таблица, кажется, открывает безграничные возможности. Действительно, если смотреть на S_0, S_1 и S_2 не как на потенциальные результаты, а как на обычные переменные, у нас есть десятки методов интерполяции для заполнения пробелов или, как сказали бы статистики, условного расчета недостающих данных некоторым оптимальным образом.

Таблица 12. Вымышленные данные для примера с потенциальными результатами

Сотрудник (u)	EX (u)	ED (u)	S_0 (u), долл.	S_1 (u), долл.	S_2 (u), долл.
Элис	6	0	81 000	?	?
Берт	9	1	?	92 500	?
Кэролайн	9	2	?	?	97 000
Дэвид	8	1	?	91 000	?
Эрнест	12	1	?	100 000	?
Фрэнсис	13	0	97 000	?	?

Один из распространенных подходов — сопоставление. Мы ищем пары людей, которые хорошо совпадают по всем переменным, кроме интересующей нас, а затем заполняем их строки, чтобы они соответствовали друг другу. Явный пример здесь — случай Берта и Кэролайн, которые идеально совпадают по стажу. Мы предполагаем, что, если бы у Берта была магистерская степень, он получал бы столько же, сколько Кэролайн (97,0 тысяч долларов), а если бы у Кэролайн была только степень бакалавра, она получал бы, как Берт (92,5 тысяч долларов). Обратите внимание, что сопоставление подразумевает ту же идею, что и ограничение по какому-то фактору (или расслоение): мы выбираем для группы, которые разделяют наблюдаемую характеристику, и используем сравнение, чтобы сделать вывод о характеристиках, которые у них, похоже, не совпадают.

Зарплату Элис трудно оценить таким образом, потому что в данных, которые я привел, для нее нет совпадения. Тем не менее статистики разработали весьма тонкие методы, чтобы сделать условный расчет на основе приблизительных совпадений, и Рубин был одним из пионеров этого подхода. К сожалению, даже самые одаренные его представители не могут превратить данные в потенциальные результаты — даже приблизительно. Ниже я покажу, что правильный ответ принципиально зависит от того, влияет образование на опыт или наоборот, о чем в таблице нет никакой информации.

Второй возможный метод — это линейная регрессия (не путать со структурными уравнениями). В этом подходе мы делаем вид, что данные пришли из какого-то неизвестного случайного источника, и используем стандартные статистические методы, чтобы найти линию (или в данном случае плоскость), которая наилучшим образом соответствует данным. Результатом такого подхода выступает уравнение, которое выглядит следующим образом:

$$S = \$65\,000 + 2\,500\,¥\,EX + 5\,000\,¥\,ED \qquad (4)$$

Уравнение (4) говорит нам, что базовая зарплата сотрудника без опыта и только с аттестатом об окончании средней школы

составляет (в среднем) 65,0 тысяч долларов. За каждый год опыта заработная плата увеличивается на 2,5 тысяч, а за каждую дополнительную образовательную ступень (до двух) зарплата увеличивается на 5,0 тысяч долларов. Соответственно, аналитик регрессии заявил бы, что наша оценка заработной платы Элис, если бы та имела высшее образование, составляла $\$65\,000 + \$2\,500 ¥ 6 + \$5\,000 ¥ 1 = \$85\,000$.

Простота и привычность таких методов объясняет, почему представление Рубина о причинном выводе как о проблеме отсутствия данных пользуется популярностью. Увы, какими бы безобидными ни казались эти методы интерполяции, они в корне ошибочны. Они основаны на данных, а не на модели. Все недостающие сведения заполняются путем изучения других значений в таблице. Как мы узнали благодаря Лестнице Причинности, любой такой метод обречен с самого начала; никакие методы, основанные лишь на данных (первый уровень), не могут ответить на контрфактивные вопросы (третий уровень).

Прежде чем сравнить эти методы со структурной каузальной моделью, давайте исследуем, почему условный расчет без учета модели не работает. В частности, объясним, почему Берт и Кэролайн, которые идеально соответствуют друг другу в плане опыта, на самом деле могут быть совершенно несравнимы, когда дело дойдет до потенциальных результатов. Еще удивительнее, что рациональная причинно-следственная история (подходящая для табл. 12) показала бы: наибольшее соответствие по зарплате у Кэролайн будет с тем, кто не соответствует ей по стажу.

Для начала нужно понять, что стаж, скорее всего, будет зависеть от образования. В конце концов, сотрудникам, получившим диплом, потребовалось для этого четыре года жизни. Таким образом, если бы у Кэролайн была только одна ступень образования (как у Берта), она могла бы использовать это дополнительное время, чтобы получить больший стаж. В этом случае у нее было бы такое же образование, но стаж солиднее, чем у Берта. Таким образом, мы можем заключить, что S_1 (Кэролайн) $> S_1$ (Берт) вопреки тому, что предсказы-

вало бы наивное сопоставление. Мы видим, что, если у нас есть причинно-следственная история, в которой образование влияет на стаж, сопоставление на основе последнего приведет к несоответствию в потенциальной зарплате.

Удивительно, но равный стаж, который вначале выглядел как приглашение к поиску соответствий, теперь превратился в громкое предупреждение против него. Табл. 12, конечно же, продолжит молчать о таких опасностях. По этой причине я не разделяю стремление Холланда рассматривать причинный вывод как проблему отсутствия данных. Наоборот. Недавняя работа Картики Мохан, моей бывшей студентки, показывает, что даже стандартные задачи с отсутствующими данными нуждаются в причинно-следственном моделировании для их решения.

Теперь давайте посмотрим, как те же данные будут обработаны с помощью структурной причинно-следственной модели. В первую очередь, прежде чем даже посмотрим на данные, нарисуем диаграмму причинности (рис. 53). На ней представим причинно-следственную историю, стоящую за данными, согласно ей стаж «слушает» образование, а зарплата — и то и другое. Фактически мы определили важные вещи, просто взглянув на диаграмму. Если бы наша модель была неправильной и *EX* было бы причиной *ED*, а не наоборот, то стаж был бы конфаундером и подбор сотрудников с аналогичным опытом был бы полностью уместным. С *ED* как причиной *EX* стаж выступает в роли посредника. Как вы уже наверняка знаете, если перепутать медиатор с конфаундером, мы совершим один из самых страшных грехов в области причинного вывода, что приведет к вопиющим ошибкам. Конфаундер нуждается в поправке, медиатор ее не допускает.

До этого момента в книге я использовал весьма неформальное слово «слушание», чтобы показать, что я имею в виду под стрелками на диаграмме причинности. Но теперь пришло время добавить немного математической плоти к этой концепции. Именно здесь структурные причинно-следственные модели отличаются от байесовских сетей или регрессионных моделей. Когда я говорю, что зарплата слушает образование

и стаж, я имею в виду, что такова математическая функция этих переменных: $S = f_S(EX, ED)$. Но нам нужно учитывать индивидуальные вариации, поэтому мы расширяем эту функцию и записываем ее как $S = f_S(EX, ED, U_S)$, где U_S означает ненаблюдаемые переменные, которые влияют на заработную плату. Мы знаем, что эти переменные существуют (например, Элис дружит с президентом компании), но они слишком разнообразны и многочисленны, чтобы явно включить их в нашу модель.

Рис. 53. Диаграмма причинности, показывающая эффект воздействия образования (*ED*) и стажа (*EX*) на зарплату (*S*)

Давайте посмотрим, как это отразится на нашем примере образования / стажа / заработной платы, предполагая во всем линейные функции. Мы используем те же статистические методы, что и раньше, с целью найти наиболее подходящее линейное уравнение. Результат будет выглядеть так же, как уравнение (4), но с одним небольшим отличием:

$$S = \$65\,000 + 2\,500\,¥\,EX + 5\,000\,¥\,ED + U_S \qquad (5)$$

Однако формальное сходство между уравнениями (4) и (5) глубоко обманчиво; их интерпретации различаются как день и ночь. Тот факт, что мы решили регрессировать *S* по *ED* и *EX* в уравнении (4), никоим образом не означает, что S слушает *ED* и *EX* в реальном мире. Это был исключительно наш выбор, и наши данные никак не помешали бы нам регрессировать *EX* по *ED* и *S* или следовать любому другому порядку. (вспомните открытие Фрэнсиса Гальтона, описанное в главе 2, о том, что регрессия не видит причины). Мы теряем эту свободу, когда

объявляем уравнение структурным. Другими словами, автор уравнения (5) должен взять на себя обязательство составлять выражения, отражающие его представления о том, кто кого слушает в реальном мире. В нашем случае он считает, что *S* действительно слушает *EX* и *ED*. Что еще более важно, отсутствие уравнения $ED = f_{ED}(EX, S, U_{ED})$ в модели означает, что *ED* предположительно не учитывает изменения в E*X* или *S*. Это различие в обязательствах дает структурным уравнениям возможность поддерживать контрфактивность, что нереально для уравнений регрессии.

В соответствии с рис. 53 у нас также должно быть структурное уравнение для *EX*, но теперь мы установим коэффициент при *S* как равный нулю, чтобы отразить отсутствие стрелки от S к *EX*. После того как мы оценим коэффициенты на основе имеющихся данных, уравнение будет выглядеть примерно так:

$$EX = 10 - 4ED + U_{EX} \quad (6)$$

Это уравнение говорит о том, что средний стаж для людей без степени магистра составляет десять лет и что каждая ступень образования (до двух) снижает *EX* в среднем на четыре года. Кроме того, обратите внимание на ключевое различие между структурными уравнениями и уравнениями регрессии: переменная *S* не входит в уравнение (6), несмотря на то, что *S* и *EX*, вероятно, сильно коррелированы. Это отражает уверенность аналитика в том, что на стаж *EX*, приобретенный любым человеком, никак не влияет его текущая зарплата.

Теперь давайте продемонстрируем, как выводить контрфактивные суждения из структурной модели. Чтобы оценить зарплату Элис, если бы у нее было высшее образование, мы сделаем три шага.

1. Абдукция: используйте данные об Элис и других сотрудниках, чтобы оценить ее специфические факторы: U_S (Элис) и U_{EX} (Элис).
2. Действие: используйте оператор *do*, меняя модель так, чтобы она отражала контрфактивное допущение — в данном случае о наличии у нее высшего образования: *ED* (Элис) = 1.

3. Прогноз: рассчитайте новую зарплату Элис, используя модифицированную модель и обновленную информацию об экзогенных переменных: U_S (Элис), U_{EX} (Элис) и *ED* (Элис). Эта рассчитанная заново зарплата равна $S_{ED} = 1$ (Элис).

Для шага 1 мы извлекаем из наших данных сведения, что *EX* (Элис) = 6 и *ED* (Элис) = 0. Мы подставляем эти значения в уравнения (5) и (6). Затем уравнения сообщают нам специфические для Элис факторы: U_S (Элис) = \$1 000 и U_{EX} (Элис) = –4. Они представляют все уникальное, особенное и чудесное, что есть в Элис. Что бы это ни было, оно добавляет 1 000 долларов к ее прогнозируемой зарплате.

Шаг 2 велит нам использовать *do*-оператор, чтобы стереть стрелки, указывающие на переменную, для которой установлено контрфактивное значение (образование), и присвоить Элис диплом бакалавра (*образование* = 1). В этом примере шаг 2 тривиален, потому что нет стрелок, указывающих на образование, и, следовательно, нет стрелок, которые нужно стереть. Однако в более сложных моделях удаление стрелок нельзя пропустить, потому что оно влияет на вычисления в шаге 3. Переменным, которые могли повлиять на результат через промежуточную переменную, больше не разрешается это делать.

Наконец, шаг 3 предполагает обновление модели с целью отразить новую информацию: U_S = \$1 000, U_{EX} = –4 и *ED* = 1. Сначала мы используем уравнение (6), чтобы пересчитать, каким был бы стаж Элис, если бы она училась в колледже: EX_{ED} = 1 (Элис) = 10 – 4 – 4 = 2 года. Затем мы используем уравнение (5), чтобы пересчитать ее потенциальную зарплату:

$$S_{ED} = 1 \text{ (Элис)} = \$65\,000 + 2\,500 ¥ 2 + \\ + 5\,000 ¥ 1 + 1\,000 = \$76\,000.$$

Наш результат S_1 (Элис) = \$76 000 — это действительная оценка потенциальной зарплаты Элис; т.е. совпадение возможно, если допущения модели верны. Поскольку в примере используется очень простая причинно-следственная модель

и элементарные (линейные) функции, различия между ней и методом регрессии на основе данных могут показаться незначительными. Но незначительные различия на поверхности отражают огромные различия в глубине. Какой бы контрфактивный (потенциальный) результат мы ни получили от структурного метода, он логически следует из допущений, отраженных в модели. В то же время результат, полученный с помощью метода, основанного на данных, будет так же своеобразен, как и ложные корреляции, поскольку он оставляет эти допущения неучтенными.

Этот пример заставил нас углубиться в тонкости причинно-следственных моделей, сильнее, чем где-либо выше в этой книге. Но позвольте мне сделать небольшое отступление и порадоваться чуду, которое стало возможным благодаря случаю с Элис. Используя комбинацию данных и модели, мы смогли предсказать поведение индивида (Элис) в полностью гипотетических условиях. Конечно, бесплатного сыра не бывает: мы получили такие веские результаты, потому что сделали веские допущения. Мы не только утвердили причинно-следственные связи между наблюдаемыми переменными, но и предположили, что функциональные связи были линейными. Но линейность здесь не так важна, как знание этих конкретных функций. Они позволили нам вычислить специфические особенности Элис по ее наблюдаемым характеристикам и обновить модель, как того требует трехэтапная процедура.

Рискуя несколько омрачить нашу радость, я должен сказать, что эта функциональная информация не всегда будет доступна на практике. В целом мы называем модель полностью заданной, если функции, выраженные стрелками, известны, и частично заданной — в иных случаях. Например, как и в байесовских сетях, мы можем знать только вероятностные отношения между родителями и детьми. Если модель задана частично, мы не оценим точно зарплату Элис; вместо этого нам, скорее всего, придется сделать утверждение с вероятностным интервалом, предположим: «Вероятность того, что ее зарплата составит 76 000 долларов, составляет 10—20%». Но даже таких вероятностных ответов достаточно для многих случаев. Более

того, действительно поражает, сколько информации мы в состоянии извлечь из диаграммы причинности, даже если у нас нет сведений о конкретных функциях, скрытых за стрелками, или есть лишь очень общие данные, скажем предположение о монотонности, с которым мы столкнулись в последней главе.

Шаги с 1 по 3, описанные выше, можно суммировать в первом законе причинного вывода, как я его называю: $Y_X(u) = Y_{MX}(u)$. Это то же самое правило, которое мы использовали в примере с расстрельной командой в главе 1, только функции здесь другие. Первый закон гласит, что потенциальный результат $Y_X(u)$ можно условно исчислить, перейдя к модели M_X (с удаленными стрелками к X) и вычислив в ней результат $Y(u)$. Отсюда следуют все оцениваемые величины на второй и третьей ступенях Лестницы Причинности. Короче говоря, сведение контрфактивных суждений к алгоритму позволяет нам завоевать столько территории на третьем уровне, сколько позволит математика, но, конечно, ни на йоту не больше.

О том, как важно видеть собственные допущения

Метод SCM, который я показал для вычисления контрфактивов, — не тот метод, который использовал бы Рубин. Основное различие между нами — применение диаграмм причинности. Они позволяют исследователям представить причинные допущения в терминах, которые они могут понять, а затем рассмотреть все контрфактивные утверждения как производные свойства от их модели мира. Причинная модель Рубина рассматривает контрфактивы как абстрактные математические объекты, которыми управляет алгебраический аппарат, а не производные от модели.

В отсутствие графического представления пользователь причинной модели Рубина обычно должен принять допущения. Первое из них, допущение о стабильном эффекте воздействия на единицу, достаточно прозрачно. В нем говорится, что каждый индивид (или единица — предпочтительный термин среди

разработчиков причинных моделей) получит одинаковый эффект от лечения независимо от того, какое лечение получают другие индивиды (или единицы). Во многих случаях, если не считать эпидемии и другие коллективные взаимодействия, это имеет смысл. Например, если предположить, что головная боль не заразна, моя реакция на аспирин не будет зависеть от того, получит ли аспирин Джо.

Второе допущение в модели Рубина, тоже безобидное, называется постоянством. Оно подразумевает, что человек, который принял аспирин и выздоровел, также выздоровеет, если получит аспирин в экспериментальном порядке. Это разумное предположение, которое рассматривается как теорема в рамках SCM, фактически утверждает, что эксперимент лишен эффекта плацебо и других недостатков.

Но главное допущение, которое неизменно должны делать все, кто использует потенциальные результаты, называется игнорируемостью. Это более технический аспект, но он является критически важной частью всей операции, поскольку аналогичен условию обмениваемости у Джейми Робинса и Сандера Гренланда, которое обсуждается в главе 4. «Игнорируемость» выражает то же требование в терминах переменной потенциального результата Y_X Она требует, чтобы Y_X не зависел от фактически полученного лечения, т.е. X, с учетом значений определенного набора конфаундеров (или конфаундеров Z. Прежде чем исследовать ее интерпретацию, мы должны признать, что любое допущение, выраженное как условная независимость, наследует широкий набор знакомых математических механизмов, разработанных статистиками для обычных (не контрфактивных) переменных. Например, статистики обычно используют правила для определения, когда одна условная независимость следует из другой. К чести Рубина, он признал, что перевод причинного понятия неосложненности в синтаксис теории вероятностей имеет смысл, пусть и на контрфактивных переменных. Допущение игнорируемости делает причинную модель Рубина действительной моделью. Табл. 12 сама по себе не является моделью, поскольку не содержит допущений о мире.

К сожалению, я еще не нашел ни одного человека, который мог бы объяснить, что такое игнорируемость на языке, на котором говорят те, кому необходимо сделать это допущение или оценить его правдоподобие для конкретной задачи. Вот до чего удалось додуматься мне. Определение пациентов в экспериментальную или контрольную группы игнорируется, если в любой страте осложнителя Z пациенты, у которых может быть один потенциальный результат — $Y_X = y$, могут оказаться в экспериментальной или контрольной группе с той же вероятностью, что и пациенты, у которых может быть другой потенциальный результат — $Y_X = Y¢$. Это определение вполне оправданно, если у вас есть функция вероятности. Но как биолог или экономист, обладающий только научными знаниями, должен оценить, правда это или нет? И как ученый оценит, сохраняется ли игнорируемость для любого из примеров, обсуждаемых в этой книге?

Чтобы понять, в чем здесь сложность, попробуем применить это объяснение к нашему примеру. Чтобы определить, является ли *ED* игнорируемым (при условии *EX*), мы должны оценить, будут ли сотрудники с одной потенциальной зарплатой, скажем $S_1 = s$, иметь одинаковый уровень образования с той же вероятностью, что и сотрудники с другой потенциальной зарплатой, предположим $S_1 = s¢$. Если вы думаете, что это похоже на замкнутый круг, я могу только согласиться! Мы хотим определить потенциальную зарплату Элис, но, еще не начав — еще не получив намека на ответ, — должны размышлять о том, зависит ли результат от *ED* или нет в каждой страте *EX*. Это настоящий когнитивный кошмар!

Как оказалось, *ED* в нашем примере нельзя игнорировать по отношению к *S*, обусловленной *EX*, и поэтому метод сопоставления (приравнивание Берта к Кэролайн) даст неправильный ответ. Фактически оценки для них должны отличаться на сумму S_1 (Берт) — S_1 (Кэролайн) = \$5 000 (читатель теперь выведет это из чисел в табл. 12 и трехэтапной процедуры). Теперь я покажу, что благодаря диаграмме причинности студент мог сразу увидеть, что *ED* нельзя игнорировать и пытаться искать здесь соответствие. При отсутствии диаграммы у сту-

дента возникнет соблазн предположить, что игнорируемость сохраняется по умолчанию, и он попадет в эту ловушку. (Это не безосновательные подозрения. Я позаимствовал идею для примера из статьи в «Юридическом журнале Гарвардского университета», где история была, по сути, такой же, как на рис. 53 и автор действительно использовал сопоставление.)

Вот как целесообразно использовать причинно-следственную диаграмму для проверки (условной) игнорируемости. Чтобы определить, является ли X игнорируемым относительно результата Y, обусловленного набором совпадающих переменных Z, нам надо лишь убедиться, что Z блокирует все обходные пути между X и Y и ни один член Z не является потомком X. Это так просто! В нашем примере предлагаемая совпадающая переменная (стаж) блокирует все лазейки (потому что их нет), но не проходит тест, так как является потомком образования. Следовательно, ED нельзя игнорировать, а EX нельзя использовать для сопоставления. Никакой сложной умственной гимнастики не требуется, достаточно взглянуть на схему. От исследователя вообще не требуется мысленно оценивать, насколько вероятен потенциальный результат того или иного лечения.

К сожалению, Рубин не рассматривает диаграммы причинности как средство, которое помогает сделать причинно-следственные выводы. Следовательно, те, кто последует его советам, не смогут проверить игнорируемость вот так. Им придется либо заняться сложной умственной гимнастикой, чтобы убедиться в верности допущения, либо просто принять его как «черный ящик». И действительно, видный исследователь потенциальных результатов Маршалл Джоффе писал в 2010 году, что допущения об игнорируемости обычно делают потому, что они оправдывают использование доступных статистических методов, а не потому, что искренне в них верят.

С прозрачностью тесно связано понятие проверяемости, которое неоднократно упоминалось в этой книге. Модель, представленная как диаграмма причинности, легко тестируется на совместимость с данными, тогда как модель, представленная на языке потенциальных результатов, лишена этой возможности. Проверка проходит так: всякий раз, когда все пути между X

и Y на диаграмме блокируются набором узлов Z, в данных X и Y должны быть независимыми при условии Z. Это свойство d-сепарации, упомянутой в главе 7, которое позволяет нам отклонять модель всякий раз, когда независимость не проявляется в данных. Напротив, если одна и та же модель выражается на языке потенциальных результатов (т.е. в виде набора утверждений об игнорируемости), нам не хватает математического аппарата, чтобы выявить независимость, которую влечет за собой эта модель, и исследователям не удастся подвергнуть ее проверке. Трудно понять, как исследователям потенциальных результатов удавалось мириться с этим недостатком, не сопротивляясь. У меня есть единственное объяснение: их так долго держали в стороне от графических инструментов, что они забыли, что каузальные модели могут и должны быть проверены.

Теперь я должен применить те же стандарты прозрачности к себе и расказать немного больше о допущениях, воплощенных в структурной модели причинности.

Помните историю Авраама, которую я упоминал выше? Первой реакцией библейского героя на известие о неминуемом разрушении Содома были поиски зависимости «доза — реакция», или функции-ответа, связывающей порочность города с его наказанием. Это был здравый научный инстинкт, но, подозреваю, мало кто из нас был бы достаточно спокоен, чтобы отреагировать таким образом.

Функция-ответ — ключевая составляющая, которая позволяет СМП обрабатывать контрфактивы. Она подразумевается в парадигме потенциальных результатов у Рубина, но является основным отличием СМП от байесовских сетей, включая каузальные байесовские сети. В вероятностной байесовской сети стрелки к Y означают, что вероятность Y определяется таблицами условной вероятности для Y с учетом наблюдений за его родительскими переменными. То же верно и для каузальных байесовских сетей, только в таблицах условной вероятности указывается вероятность Y с учетом интервенций по родительским переменным. Обе модели определяют вероятности для Y, а не конкретное значение Y. В структурной модели причинности нет дополнительных таблиц вероятностей. Стрелки просто

означают, что Y является функцией от своих родителей, так же как и экзогенная переменная U_Y:

$$Y = f_Y(X, A, B, C, \ldots, U_Y) \qquad (8.4)$$

Таким образом, инстинкт Авраама был здравым. Чтобы превратить некаузальную байесовскую сеть в причинную модель, или, точнее, сделать ее способной отвечать на контрфактивные запросы, нам нужна взаимосвязь «доза — реакция» в каждом узле.

Я осознал это далеко не сразу. Еще не обратившись к контрфактивам, я очень долго пытался сформулировать модели причинности, используя таблицы условной вероятности. Одним из препятствий, с которым я столкнулся, были циклические модели, полностью устойчивые к формулировкам условной вероятности. Еще одним препятствием была необходимость придумать запись, позволяющую отличать вероятностные байесовские сети от причинных. В 1991 году меня внезапно осенило, что все трудности исчезнут, если сделать Y функцией от его родительских переменных и обозначить с помощью U_Y все неопределенности, касающиеся Y. В то время это казалось ересью по отношению к моему же учению. Посвятив несколько лет изучению причин вероятностей в искусственном интеллекте, я предлагал теперь сделать шаг назад и использовать невероятностную квазидетерминированную модель. Я до сих пор помню, как мой тогдашний студент Дэнни Гейгер недоверчиво спрашивал: «Детерминированные уравнения? Действительно детерминированные?» Как будто Стив Джобс только что велел ему купить PC вместо Mac. (Это был 1990 год!)

На первый взгляд, в этих уравнениях не было ничего революционного. Экономисты и социологи использовали такие модели с 1950—60-х годов и называли это моделированием структурных уравнений. Но это название сигнализирует о противоречиях и путанице, связанной с каузальной интерпретацией уравнений. Со временем экономисты упустили из виду тот факт, что первые разработчики этих моделей, Трюгве Ховельмо в экономике и Отис Дадли Дункан в социологии, хотели, чтобы

они отображали причинно-следственные связи. Они начали путать структурные уравнения с линиями регрессии, тем самым отрывая суть от формы. Например, в 1988 году, когда Дэвид Фридман попросил 11 исследователей SEM объяснить, как применять интервенцию к модели структурного уравнения, ни один из них не смог этого сделать. Они рассказали, как оценить коэффициенты на основе данных, но не сумели растолковать, зачем это делать. Если интерпретация функции-ответа, которую я представил в период с 1990 по 1994 год, и внесла нечто новое, то это было всего лишь возвращением и оформлением изначальных намерений Ховельмо и Дункана. Я хотел представить их ученикам смелые выводы, которые вытекают из этих намерений, если относиться к ним серьезно.

Некоторые из этих выводов поразили бы даже Ховельмо и Дункана. Возьмем, к примеру, идею о том, что из каждого SEM, сколько угодно простого, можно вывести все контрфактивы, какие только получается вообразить среди переменных в модели. Наша способность вычислить потенциальную зарплату Элис, если бы она имела высшее образование, вытекала из этой идеи. Но даже сегодня современные экономисты все еще не усвоили эту идею.

Еще одно важное различие между SEM и SCM, помимо средней буквы, заключается в том, что взаимосвязь между причинами и следствиями в SCM не обязательно линейна. Методы, вытекающие из анализа SCM, действительны как для нелинейных, так и для линейных функций и как для дискретных, так и для непрерывных переменных.

У линейных SEM есть много преимуществ и много недостатков. С точки зрения методологии, они соблазнительно просты. Их легко оценить на основе наблюдений с помощью линейной регрессии, на что способны десятки статистических программ, которые сделают это за вас.

Однако линейные модели не способны представлять кривые «доза — эффект», которые не являются прямыми линиями. Они не в состоянии отражать пороговые эффекты, например, для лекарства, действие которого усиливается до определенной дозы, а потом прекращается. Они также не представляют

взаимодействия между переменными. Так, линейная модель не опишет ситуацию, в которой одна переменная усиливает или подавляет эффект другой (предположим, образование может усилить эффект стажа, поскольку позволит получить работу с более быстрым карьерным продвижением и более высокими ежегодными прибавками).

Несмотря на то что споры о верных допущениях неизбежны, наша основная идея довольно проста: радуйтесь! Благодаря полностью определенной SCM, включающей диаграмму причинности и все стоящие за ней функции, мы способны ответить на любой контрфактивный вопрос. Даже с частичной SCM, где некоторые переменные скрыты или отношения «доза — эффект» неизвестны, мы все же можем во многих случаях ответить на поставленный вопрос. В следующих двух разделах приведены некоторые примеры.

Контрфактивные суждения и закон

Теоретически, контрфактивы должны с легкостью использоваться в зале суда. Я говорю «теоретически», потому что юристы очень консервативны. Им требуется много времени, чтобы принять новые математические методы. Но использование контрфактивных суждений фактов в качестве аргументов на самом деле известно в юридической практике очень давно как «то, без чего невозможно».

«Примерный уголовный кодекс» США формулирует это следующим образом: «Поступок является причиной результата, когда: (а) он предшествует данному результату таким образом, что без него результат не наступил бы». Если обвиняемый выстрелил из пистолета и пуля попала в жертву и убила ее, стрельба из пистолета является необходимой причиной смерти (без которой смерть не наступила бы), поскольку жертва была бы жива, если бы не стрельба. Но причины также могут быть косвенными. Если Джо завалил доступ к пожарной лестнице мебелью, а Джуди гибнет на пожаре, не сумев выбраться

наружу, то Джо несет юридическую ответственность за ее смерть, даже если он не разводил огонь.

Как выразить необходимые причины в терминах потенциальных результатов? Если мы допустим, что результатом Y будет «смерть Джуди» (с $Y = 0$, если Джуди жива, и $Y = 1$, если Джуди умирает), а эффектом X будет «заблокированная Джо пожарная лестница» (с $X = 0$, если Джо ее не блокировал, и $X = 1$, если он это сделал), то предлагается задать следующий вопрос: учитывая, что пожарная лестница действительно была заблокирована ($X = 1$) и Джуди умерла ($Y = 1$), какова вероятность того, что Джуди выжила бы ($Y = 0$), если бы X был равен 0?

Символически вероятность, которую мы хотим оценить, выглядит как P ($Y_{X=0} = 0 \mid X = 1, Y = 1$). Поскольку эта формула довольно громоздкая, я сокращу ее как PN (*the Probability of Necessity* — «вероятность необходимости», т.е. вероятность того, что $X = 1$ является необходимой или непредвиденной причиной $Y = 1$).

Заметим, что вероятность необходимости включает контраст между двумя разными мирами: реальным миром, где $X = 1$, и контрфактивным миром, где $X = 0$ (выражается индексом $X = 0$). Фактически ретроспективный взгляд (знание того, что произошло в реальном мире) — это критическое различие между контрфактивами (третий уровень Лестницы Причинности) и интервенцией (второй уровень). Без ретроспективного взгляда нет никакой разницы между P ($Y_{X=0} = 0$) и P ($Y = 0 \mid do$ ($X = 0$)). В обоих случаях выражена вероятность того, что в нормальных условиях Джуди будет жива, если мы гарантируем, что выход не заблокирован; они не включают пожар, смерть Джуди или заблокированный выход. Но ретроспективный взгляд способен изменить нашу оценку вероятностей. Предположим, мы заметили, что $X = 1$ и $Y = 1$ (ретроспективно). Тогда $З$ ($Н_{Ч=0} = 0$ / $Ч = 1$б $Н = 1$)) не равно $З$ ($Н_{Ч=0} = 0$ / $Ч = 1$). Знание того, что Джуди умерла ($Y = 1$), дает нам информацию об обстоятельствах, которую мы не получили бы, просто зная, что дверь была заблокирована ($X = 1$). По крайней мере, мы оценим серьезность пожара.

Фактически можно показать, что не существует способа отразить $P\,(Y_{X=0} = 0 \mid X = 1, Y = 1)$ в *do*-выражении. Хотя это может показаться довольно туманным, на деле мы имеем математическое доказательство того, что контрфактивы (третий уровень) находятся выше интервенции (второй уровень) на Лестнице Причинности.

В последних нескольких абзацах мы почти незаметно ввели в обсуждение понятие вероятности. Юристы давно поняли, что математическая определенность — слишком высокий стандарт для доказательств. Верховный суд США в 1880 году постановил относительно уголовных дел, что вина должна быть доказана «вне разумных сомнений». Не вне всяких сомнений, а вне разумных сомнений. Верховный суд никогда не давал точного определения для этого термина, но предполагается, что существует некоторый порог, возможно 99 или 99,9% вероятности, выше которого сомнение становится необоснованным и в интересах общества заключить обвиняемого под стражу. В гражданском (но не в уголовном) судопроизводстве стандарт для доказательств несколько яснее. Закон требует «преобладания доказательств», что обвиняемый причинил вред, и представляется разумным интерпретировать это как вероятность, превышающую 50%.

Хотя принцип необходимой причины в целом принят, юристы признают, что в некоторых случаях он приводит к судебной ошибке. Одним из классических примеров является сценарий падающего пианино, когда обвиняемый стреляет в жертву и промахивается, но в процессе бегства с места происшествия жертва попадает под падающее пианино и погибает. После проверке этим принципом обвиняемого надо было бы признать виновным в убийстве, потому что потерпевший не оказался бы рядом с падающим пианино, если бы не убегал. Но интуиция подсказывает нам, что подсудимый не виновен в убийстве (хотя может быть виновен в покушении на убийство), потому что никак не мог предвидеть падения пианино. Адвокат сказал бы, что *непосредственной причиной* смерти является пианино, а не выстрел.

Понятие непосредственной причины гораздо более туманно. «Примерный уголовный кодекс» гласит, что результат не должен быть «слишком отдаленным или случайным, чтобы иметь [обоснованное] отношение к ответственности исполнителя или тяжести его преступления». В настоящее время задача определить это остается на усмотрение судьи. Я бы предположил, что это форма *достаточной причины*: были ли действия подсудимого достаточными, чтобы с достаточно высокой вероятностью вызвать событие, которое на самом деле привело к смерти?

Хотя смысл непосредственной причины весьма расплывчат, смысл достаточной причины довольно ясен. Используя контрфактивные обозначения, мы можем определить вероятность достаточности (*the Probability of Sufficiency; PS*) как равную $P\,(Y_{X=1} = 1 \mid X = 0, Y = 0)$. Эта запись заставляет нас представить ситуацию, когда $X = 0$ и $Y = 0$: обвиняемый не стрелял в жертву и жертва не бежала под летящее пианино. Затем мы спрашиваем: насколько вероятно, что в такой ситуации выстрел ($X = 1$) приведет к результату $Y = 1$ (попадание под пианино)? Этот вопрос требует контрфактивного суждения, но, думаю, большинство из нас согласятся, что вероятность такого исхода будет чрезвычайно мала. И наша интуиция, и «Примерный уголовный кодекс» предполагают, что, если *PS* слишком мала, мы не должны обвинять ответчика в причинении $Y = 1$.

Поскольку разница между необходимыми и достаточными причинами очень важна, я думаю, стоит закрепить эти две концепции на простых примерах. Достаточная причина встречается чаще, и мы уже встречались с этой концепцией в случае с расстрельной командой в главе 1. Там было достаточно, чтобы выстрелил либо солдат A, либо солдат *B*, чтобы вызвать смерть заключенного, но ни то ни другое (само по себе) не было необходимо. Итак, $PS = 1$ и $PN = 0$.

Все становится немного интереснее, когда наступает неопределенность, например, если каждый солдат имеет вероятность не подчиниться приказам или промахнуться. Предположим, у солдата A есть вероятность не попасть p_A, тогда его *PS* будет $1 — p_A$, ибо это его вероятность поразить цель и вызвать смерть.

Его *PN*, однако, будет зависеть от того, насколько вероятно, что солдат *B* воздержится от стрельбы или промахнется. Только при таких обстоятельствах у солдата *A* возникла бы необходимость стрелять; т.е. заключенный был бы жив, если бы солдат *A* не выстрелил.

Классический пример, демонстрирующий необходимую причинность, — случай с пожаром, который возник, потому что кто-то зажег спичку. Задается вопрос: что вызвало возгорание, зажженная спичка или присутствие кислорода в комнате? Обратите внимание, что оба фактора одинаково необходимы, ведь без одного из них пожар не произошел бы. Итак, с чисто логической точки зрения, оба фактора в равной степени ответственны за возгорание. Почему же тогда мы считаем зажигание спички более разумным объяснением возгорания, чем присутствие кислорода?

Чтобы ответить на этот вопрос, рассмотрите два предложения:

1. Дом остался бы цел, если бы не зажгли спичку.
2. Дом остался бы цел, если бы не было кислорода.

Оба предложения верны. Тем не менее я уверен, что подавляющее большинство читателей выбрало бы первый сценарий, если бы их попросили объяснить, что стало причиной пожара — спичка или кислород. Итак, чем объясняется разница?

Ответ явно имеет некое отношение к нормальности: наличие кислорода в доме вполне нормально, что вряд ли справедливо в отношении зажженной спички. Разница не проявляется в логике, но видна в двух измерениях, которые мы обсуждали выше, *PS* и *PN*.

Если мы примем во внимание, что вероятность зажечь спичку намного ниже, чем вероятность иметь в доме кислород, мы обнаружим количественно, что для спичек высоки и *PN*, и *PS*, а для кислорода *PN* будет высоким, а *PS* — низким. Не поэтому ли мы интуитивно виним спичку, а не кислород? Вполне вероятно, но ответ может оказаться сложнее.

В 1982 году психологи Даниэль Канеман и Амос Тверски исследовали, как люди выбирают причину («Ах, если бы»),

чтобы «отменить» в воображении нежелательный результат, и обнаружили определенные закономерности. Одна из них состоит в том, что люди чаще представляют «отмену» редкого, а не обычного события. Так, если мы «отменяем» пропущенную встречу, то с большей вероятностью скажем: «Ах, если бы поезд ушел по расписанию», чем «Ах, если бы поезд ушел раньше». Кроме того, мы склонны жалеть о собственных действиях (например, что подожгли спичку), а не о событиях вне нашего контроля. Способность оценивать *PN* и *PS* на основе нашей модели мира предполагает, что у нас есть систематический способ принимать во внимания эти соображения, а значит, когда-нибудь мы научим роботов давать содержательные объяснения для необычных событий.

Мы видели, что *PN* воплощает обоснования для критерия «необходимой причины» в юридической практике. Но надо ли учитывать *PS* в уголовном и гражданском праве? Я считаю, что это необходимо, ибо внимание к достаточности подразумевает внимание к последствиям своих действий. Человек, который зажег спичку, должен был предвидеть наличие кислорода, но обычно никто не считает, что следует откачать из дома кислород перед церемонией зажжения спички.

Какой же вес необходимым и достаточным компонентам причинно-следственной связи важно придавать закону? Философы права не обсуждали правовой статус этого вопроса, возможно, потому, что понятия *PS* и *PN* не были формализованы с такой точностью. Однако, с точки зрения ИИ, очевидно, что и *PN*, и PS нужно участвовать в формировании объяснений. У робота, которому было поручено объяснить, почему возник пожар, нет другого выбора, кроме как рассмотреть и то и другое. Сосредоточенность только на *PN* подтолкнет к необоснованному выводу о том, что поджигание спички и наличие кислорода в равной степени являются подходящими объяснениями. Робот, который будет давать такое объяснение, быстро утратит доверие владельца.

Необходимые причины, достаточные причины и климатические изменения

В августе 2003 года Западную Европу поразила ужасная жара, самая сильная за пять веков. Больше всего пострадала Франция. Правительство страны заявило, что жара стала причиной 15 тысяч смертей, причем жертвами в основном оказались одинокие пожилые люди, не имевшие дома кондиционеров. Что же было настоящей причиной их смерти: глобальное потепление или неудача оказаться не в то время и не в том месте?

До 2003 года ученые-климатологи избегали размышлений над такими вопросами. Расхожее мнение было примерно таким: хотя такой феномен чаще возникает из-за глобального потепления, невозможно приписать это конкретное его проявление предшествующим выбросам парниковых газов.

Майлз Аллен, физик из Оксфордского университета и автор приведенной выше цитаты, предложил способ добиться большего: использовать показатель, называемый долей приписываемого риска (*Fraction of Attributable Risk; FAR*), чтобы количественно оценить эффект климатических изменения. Для применения FAR надо знать два числа: p_0 — вероятность аномальной жары, подобной жаре 2003 года, до климатических изменений (например, до 1800 года), и p_1 — ее вероятность после климатических изменений. Так, если вероятность удваивается, справедливо сказать, что половина риска связана с изменениями климата. Если она утраивается, то с изменениями климата связаны две трети риска.

Поскольку FAR определяется исключительно на основе данных, у этого показателя не обязательно есть причинное значение. Но оказывается, что при двух умеренных причинно-следственных допущениях она идентична вероятности необходимости. Во-первых, мы должны допустить, что воздействие (парниковые газы) и результат (жаркая погода) не осложняют друг друга — у них нет общей причины. Это весьма рационально, ведь, насколько нам известно, на выброс парниковых газов влияем только мы сами. Во-вторых, нужно

допустить монотонность. Мы кратко обсудили это допущение в предыдущей главе; в нынешнем контексте оно означает, что воздействие никогда не дает эффекта, противоположного ожидаемому, т.е. парниковые газы никогда не защитят нас от аномальной жары.

В отсутствие осложнителей и защиты FAR поднимается с первого уровня на третий уровень Лестницы Причинности, где становится *PN*. Но Аллен не знал причинной интерпретации FAR (вероятно, она не слишком популярна у метеорологов), и это заставило его представить результаты в достаточно туманных выражениях.

Но какие же данные мы можем использовать для оценки FAR (или *PN*)? Мы наблюдали только одну погодную аномалию такого рода. Мы не в состоянии провести контролируемый эксперимент, потому что для этого нужно проконтролировать уровень углекислого газа, будто щелкая выключателем. К счастью, у климатологов есть секретное оружие: они способны провести эксперимент in silico — компьютерное моделирование.

Аллен и Стотт из британской метеорологической службы взяли на себя эту задачу и в 2004 году стали первыми учеными-климатологами, которые решились сформулировать причинно-следственную связь для отдельного погодного явления. Удалось ли это? Судите сами. Вот что они написали: «Весьма вероятно, что более половины риска температурных аномалий в Европе, превышающих пороговое значение в 1,6 °C, связано с влиянием человека».

Хотя я высоко ценю храбрость Аллена и Стотта, мне жаль, что их важная находка была похоронена в дебрях непонятного языка. Я попробую разобрать это утверждение, а затем попытаюсь понять, почему они выразили его так запутанно. Во-первых, «температурные аномалии, превышающие пороговое значение 1,6 °C» — это их способ определить результат. Они выбрали этот порог, потому что средняя температура в Европе тем летом была более чем на 1,6 °C выше нормы, чего раньше не было за всю историю наблюдений. Их выбор уравновесил конкурирующие цели: остановиться на результате, достаточно экстремальном, чтобы уловить эффект от глобального поте-

пления, но не слишком привязанном к конкретике 2003 года. Чтобы не использовать, например, среднюю температуру в августе во Франции, они взяли более широкий критерий средней температуры в Европе за все лето.

Во-вторых, что они имели в виду под «весьма вероятно» и «половиной риска»? С математической точки зрения Аллен и Стотт хотели сказать следующее: вероятность, что FAR превысила 50%, составляет 90%. Или, другими словами, есть 90%-ная вероятность того, что летняя погода, как в 2003 году, будет наступать более чем в два раза чаще при нынешнем уровне углекислого газа по сравнению с доиндустриальным уровнем. Обратите внимание, что здесь есть два уровня вероятности: мы говорим о вероятности вероятности! Неудивительно, что от таких утверждений у нас вскипает мозг и плывет перед глазами. Причина удвоенной атаки в том, что на летнюю жару влияют два вида неопределенности. Первая связана с масштабом долгосрочных климатических изменений. Ей соответствуют 90% вероятности. Но, даже если масштабы долгосрочных климатических изменений известны точно, есть неопределенность в отношении погоды в любой конкретный год. Именно эта переменчивость заложена в 50%-ную долю приписываемого риска.

Таким образом, мы должны признать, что Аллен и Стотт пытались донести сложную идею. Тем не менее в их заключении отсутствует одна вещь: причинность. Там нет даже намека на причинно-следственную связь или, быть может, только намек и есть — в туманной фразе «можно отнести к человеческому влиянию».

Теперь сравните это с причинной версией того же вывода: «Вероятно, выбросы CO_2 были необходимой причиной аномальной жары 2003 года». Какое предложение, их или наше, вы сможете вспомнить завтра? А какое могли бы объяснить соседу?

Я не эксперт по изменению климата, поэтому взял этот пример у одного из моих соратников Алексиса Ханнарта

из Франко-аргентинского института изучения климата и его эффектов в Буэнос-Айресе. Ханнарт — приверженец причинно-следственного анализа. в климатологии. Ханнарт изобразил график причинности на рис. 54. Поскольку парниковые газы — узел верхнего уровня в климатической модели и к нему не ведут стрелки, он утверждает, что по ним и климатической реакции нет никаких осложнений. Подобным образом он ручается за допущение об отсутствии защиты (т.е. парниковые газы не могут защитить нас от аномальной жары).

Ханнарт идет дальше Аллена и Стотта и использует наши формулы для вычисления вероятности достаточности (*PS*) и необходимости (*PN*). В случае с аномальной жарой в Европе в 2003 году он обнаружил, что *PS* была чрезвычайно низкой, около 0,007 2, а значит, было невозможно предсказать эту аномалию в том конкретном году. В свою очередь, вероятность необходимости PN составила 0,9, что согласуется с результатами Аллена и Стотта. Это означает, что, весьма вероятно, без парниковых газов жары не было бы.

Очевидно низкое значение *PS* следует рассматривать в более широком контексте. Мы не просто хотим знать вероятность аномальной жары в этом году; мы хотели бы знать вероятность ее повторения в течение более длительного периода времени, скажем в следующие 10 или 50 лет. С увеличением временных рамок *PN* уменьшается, поскольку в действие вступают другие возможные механизмы возникновения аномальной жары. Тем не менее *PS* увеличивается, потому что, по сути, возрастает риск самого неблагоприятного сценария. Согласно вычислениям Ханнарта, существует 80%-ная вероятность того, что климатические изменения вызовут такую же (или более сильную) жару в Европе, как в 2003 году, в течение 200-летнего периода. Возможно, это звучит не слишком пугающе, но речь идет о нынешней концентрации парниковых газов в атмосфере. В действительности уровень CO_2 будет расти и дальше, отчего увеличится *PS* и сократится промежуток времени до новой жары.

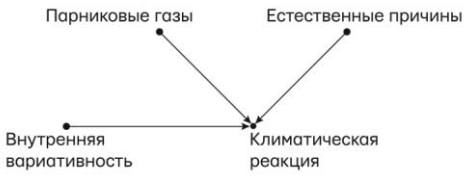


Рис. 62. Диаграмма причинности для примера с климатическими изменениями

Могут ли обычные люди научиться улавливать разницу между необходимыми и достаточными причинами? Это непростой вопрос. Даже ученые иногда сталкиваются с трудностями. Например, об аномальных температурах в России в 2010 году, когда выдалось самое жаркое лето за всю историю наблюдений и торфяные пожары омрачили небо над Москвой, вышло два исследования с противоположными выводами. Одна группа решила, что жара была вызвана естественными колебаниями температур, другая — что причиной стало изменение климата. По всей видимости, разночтения возникли из-за того, что группы по-разному определили результат. Первая, по-видимому, строила аргументы на основе *PN* и получила высокую вероятность причины в климатических изменениях. Вторая использовала *PS* и получила низкую вероятность. Вторая группа приписала жару постоянному высокому давлению над Россией — что кажется мне достаточной причиной — и обнаружила, что парниковые газы практически не связаны с этим явлением. Но любое исследование, использующее *PS* в качестве показателя за короткий период, устанавливает высокую планку для всех, кто пытается доказать причинно-следственные связи.

Прежде чем отойти от этого примера, я хотел бы еще раз прокомментировать компьютерные модели. Большинству ученых приходится усердно работать, чтобы получить контрфактивную информацию, скажем мучительно комбинируя данные наблюдательных и экспериментальных исследований. Ученые-климатологи могут легко получить контрфактивные данные из компьютерных моделей: достаточно ввести новое

значение для концентрации углекислого газа в воздухе и дать программе поработать. «Легко», конечно, здесь понятие относительное. За простой причинно-следственной схемой на рис. 54 скрывается невероятно сложная функция-ответ, заданная миллионами строк компьютерного кода, которые используются для моделирования климата.

Здесь возникает естественный вопрос: насколько мы можем доверять компьютерному моделированию? У этого вопроса есть политические нюансы, особенно здесь, в США. Однако я постараюсь дать аполитичный ответ. Я считаю, что функция-ответ в этом примере вызывает гораздо больше доверия, чем линейные модели, которые так часто встречаются в естественных и социальных науках. Линейные модели часто выбирают только по причине удобства. Для сравнения: климатические модели отражают более чем вековые исследования физиков, метеорологов и климатологов. Это усилия ученых понять процессы, которые определяют нашу погоду и климат. По любым нормальным научным стандартам климатические модели являются веским и убедительным доказательством, но с одной оговоркой. Хотя они превосходно предсказывают погоду на несколько дней вперед, они никогда не проверялись в перспективных исследованиях на протяжении веков, поэтому все еще содержат систематические ошибки, о которых мы не знаем.

Мир контрфактивного

Я надеюсь, к этому моменту уже очевидно, что контрфактивные суждения — важный инструмент познания мира и нашего воздействия на него. Хотя мы никогда не сможем пройти по обеим дорожкам, расходящимся в лесу, во многих случаях получится с достаточной уверенностью предсказать, куда они ведут.

Несомненно, разнообразие и богатство причинно-следственных запросов, которые обрабатываются с помощью причинного вывода, значительно возрастет, если мы включим

в них контрфактивные утверждения. Другой очень популярный тип запроса, который я здесь не обсуждал, называется влиянием лечения на получивших его (*the Effect of Treatment on the Treated; ETT*). Он используется, чтобы оценить, являются ли люди, имевшие доступ к лечению, теми, кто получит от него наибольшую пользу. Этот показатель гораздо лучше отражает эффективность лечения, чем средний причинный эффект (*the Average CaUsal Effect; ACE*). ACE, который вы получите в результате рандомизированного контролируемого исследования, усредняет эффективность лечения для всей группы людей. Но что, если на практике те, кого взяли в программу, не получат от лечения наибольшую пользу? Чтобы оценить общую эффективность программы, используется ETT, которое показывает, какой эффект наблюдался бы у пациентов с неудачным исходом лечения, если бы их не лечили. Это контрфактивная мера, имеющая важнейшее значение для принятия практических решений. Мой бывший ученик Илья Шпицер (ныне сотрудник Университета Джонса Хопкинса) сделал для ETT то же, что *do*-исчисление сделало для ACE, — четко показал, когда его оценивают по данным с использованием диаграммы причинности.

Несомненно, самое популярное применение контрфактивности сегодня в науке — это анализ посредничества. По этой причине я посвящаю ему отдельную главу (главу 9). Как ни странно, многие люди, особенно если они используют классические методы для анализа посредничества, не осознают, что говорят о контрфактивном эффекте.

В научном контексте медиатор, или переменная-посредник, — это переменная, которая транслирует эффект воздействия на результат. В этой книге мы видели много образцов посредничества, например *курение → смола → рак* (где смола будет посредником). Главный вопрос, представляющий интерес в таких случаях, состоит в том, приходится ли на переменную-посредник весь эффект от переменной воздействия или какая-то его часть. Мы бы представили такой эффект отдельной стрелкой, ведущей непосредственно от воздействия к результату: *курение → рак*.

Анализ посредничества позволяет отделить прямой эффект (который не проходит через посредника) от косвенного (части, которая проходит через посредника). Его важность нетрудно уловить. Если курение вызывает рак легких только из-за смол, то повышенный риск рака устранялся бы сигаретами без смол, предположим электронными. Однако, если курение вызывает рак напрямую или через другого посредника, электронные сигареты не решат проблему. В настоящее время этот медицинский вопрос остается открытым.

На данном этапе, возможно, неочевидно, что прямые и косвенные эффекты связаны с контрфактивными утверждениями. Для меня это было совершенно неочевидно! Более того, это оказалось одним из самых больших сюрпризов в моей научной карьере. В следующей главе я рассказываю об этом и привожу много вариантов для применения такого анализа на практике.

Глава 9
Опосредование: в поисках механизма действия

Не было гвоздя — подкова упала.
Подкова упала — лошадь захромала.
Лошадь захромала — командир убит.
Конница разбита —
Армия разбита, конница бежит,
Враг заходит в город, пленных не щадя,
Потому что в кузнице не было гвоздя.
Перевод С. Маршак

В обычном языке у вопроса «Почему?» есть по крайней мере две версии. Первая прямолинейна: вы видите воздействие и хотите знать причину. Ваш дедушка в больнице, и вы спрашиваете: «Почему? Откуда у него инфаркт, если он выглядел таким здоровым?»

Но есть и другая версия вопроса «Почему?», которой мы задаемся, когда стремимся лучше понять связь между известными причиной и следствием. Например, мы сделали наблюдение, что лекарственный препарат *B* предотвращает инфаркты. Или, как Джеймс Линд, узнаем, что плоды цитрусовых помогают избежать цинги. Человеческий разум неугомонен — нам надо знать больше. Вскоре мы начинаем задавать вопрос «Почему?» в его второй версии: каков механизм, благодаря которому плоды цитрусовых предотвращают цингу? Эта глава сосредоточена как раз на этой, второй версии «Почему?».

Поиск механизмов действия очень важен как для науки, так и для повседневной жизни, потому что различные механизмы требуют различных действий в случае, если обстоятельства меняются. Допустим, у нас кончились апельсины. Зная механизм их действия, мы тем не менее предотвратим цингу — нам просто понадобится другой источник витамина С. Но, если мы не знаем этого механизма, у нас может возникнуть идея попробовать использовать для этого бананы.

Ученые называют «Почему?» вот этого второго типа словом «опосредование». Вы можете увидеть в научном журнале фразу вроде «Влияние препарата *B* на риск инфаркта опосредовано его влиянием на артериальное давление». Это утверждение описывает простую каузальную модель: *препарат B → артериальное давление → риск инфаркта*. В этом случае изучаемое вещество снижает слишком высокое АД, что в свою очередь уменьшает риск инфаркта (биологи обычно используют другой символ — *A* —| *B*, когда причина А ингибирует следствие *B*, но в работах по причинности традиционно применяют обозначения *A* → B, как для позитивных, так и для негативных каузальных воздействий). Аналогично характеризуется механизм воздействия плодов цитрусовых на цингу каузальной моделью *цитрусовые → витамин С → цинга*.

Относительно медиаторов у нас возникают определенные типичные вопросы: отвечает ли медиатор за весь эффект? Действует ли препарат *B* исключительно за счет снижения артериального давления, или, возможно, задействованы также и другие механизмы? Распространенный тип медиатора в медицине — это эффект плацебо: если некое вещество приносит пользу только благодаря вере пациента в его эффективность, большинство врачей сочтет его неэффективным. Опосредование — важная концепция также и в юриспруденции. Если мы зададимся вопросом, дискриминирует ли некая компания женщин, выплачивая им меньшее жалованье, нежели мужчинам, мы поднимаем проблему опосредования. Ответ зависит от того, насколько наблюдаемая разница зарплат вытекает непосредственно из пола кандидата на рабочее место, или же она связана с полом косвенно, будучи опосредована такими

переменными-медиаторами, как квалификация, которую работодатель уже не в состоянии контролировать.

Все приведенные выше примеры вопросов требуют четкого умения различать *суммарное воздействие*, *прямое воздействие* (которое не проходит через переменную-медиатор) и *косвенные воздействия* (которые как раз проходят). В прошлом столетии ученым было сложно даже дать этим терминам точное определение. Связанные по рукам и ногам жесткими табу на слово «причинность», некоторые из них пытались определить опосредование, избегая каузальной лексики. Другие полностью отказались от анализа медиации и объявили концепцию прямого и непрямого воздействия «скорее вредной, чем полезной для ясности статистического мышления».

Мне лично опосредование тоже не далось без борьбы, однако и принесло едва ли не самый крупный успех в моей карьере, потому что сначала я ошибался, но, учась на своих ошибках, нашел неожиданное решение. Некоторое время я полагал, что непрямые воздействия лишены оперативных последствий, потому что их, в отличие от прямых эффектов, нельзя определить в терминах интервенций. Прорыв наметился тогда, когда я осознал, что для этого подойдут термины контрфактивных высказываний и что у них могут быть важные стратегические последствия. Их реально численно оценить только после того, как мы поднимемся на третью ступень Лестницы Причинности, и поэтому я поместил их в конец этой книги. Опосредование, или медиация, расцвела в новом окружении и позволила нам оценить численно, иногда из голых данных, ту часть воздействия, которая опосредуется любым выбранным путем.

Понятно, что благодаря их контрфактивным одеяниям, непрямые воздействия остаются несколько загадочными даже для признанных лидеров Революции Причинности. Я уверен, что их очевидная польза, однако, постепенно одержит победу над все еще сохраняющимся недоверием к метафизике контрфактивного. Вероятно, их можно сравнить с иррациональными и мнимыми числами: те тоже сначала вызывали у людей некоторый дискомфорт (отсюда и термин «иррациональные»),

но постепенно их полезность превратила этот дискомфорт в восторг.

Чтобы проиллюстрировать это положение, я приведу несколько примеров того, как исследователи в рамках различных научных дисциплин почерпнули полезные идеи из анализа опосредования. Один исследователь занимался проектом реформы образования под названием «Алгебра для всех», который сначала казался провальным, но позже достиг успеха. Исследованию, посвященному применению жгутов во время военных действий в Ираке и Афганистане, не удалось показать, что это давало какие-либо преимущества; аккуратный анализ опосредования объясняет, почему эти преимущества оказались скрыты. Таким образом, в последние 15 лет Революция Причинности открыла простые и ясные правила оценки того, какая часть эффекта обусловливается прямым, а какая — косвенным воздействием. Она превратила опосредование из нечетко сформулированной концепции с сомнительной надежностью в популярный и широко используемый инструмент научного анализа.

Цинга: неверный медиатор

Я хотел бы начать с поистине ужасающего исторического случая, подчеркивающего важность понимания принципов опосредования.

Один из самых ранних примеров контролируемого эксперимента — исследование цинги, проведенное капитаном Джеймсом Линдом, результаты которого опубликованы в 1747 году. Во времена, когда жил Линд, цинга была страшным заболеванием: от нее в период между 1500 и 1800 годами погибло, по осторожным оценкам, около 2 миллионов моряков. Линд доказал, настолько аккуратно, насколько это было возможно в то время, что добавление плодов цитрусовых в рацион моряков предотвращало развитие этой страшной болезни. К началу *XIX* века во всем британском флоте цинга ушла в историю, поскольку все его корабли отплывали в море с достаточным

запасом плодов цитрусовых. Учебники обычно обрывают эту историю как раз на этом месте, отмечая грандиозный триумф научной мысли.

Тем более удивительным оказывается, что это вполне предотвратимое заболевание неожиданно вернулось через 100 лет, когда британские экспедиции стали исследовать полярные области. Все участники Британской арктической экспедиции в 1875 году, экспедиции Джексона — Хармсворта в Арктику в 1894-м и в особенности двух знаменитых экспедиций Роберта Скотта в Антарктиду очень сильно пострадали от цинги.

Как это могло случиться? В двух словах: невежество и самонадеянность — в буквальном смысле убойное сочетание. К началу XX века ведущие британские врачи успели позабыть уроки предыдущего столетия. Врач в экспедиции Скотта 1903 года доктор Реджинальд Кётлиц полагал, что цингу вызывает протухшее мясо. Более того, к этому он добавлял, что «польза так называемых противоскорбутных (т.е. предотвращающих цингу продуктов, подобных соку лайма) средств иллюзорна». Для экспедиции 1911 года Скотт запасся сушеным мясом, весьма скрупулезно исследованным на предмет возможной порчи, но ни плоды, ни соки цитрусовых в паек не вошли. Вера Скотта в мнение своего врача, по всей видимости, внесла свой вклад в последовавшую затем трагедию. Все пятеро добравшихся до Южного полюса погибли, при этом двое — от какой-то болезни, которой, скорее всего, была именно цинга. Один член экспедиции повернул обратно, не дойдя до полюса, и смог вернуться живым, но с крайне тяжелым случаем цинги.

С точки зрения сегодняшнего дня совет Кётлица граничит с уголовно наказуемой врачебной некомпетентностью. Как случилось, что уроки Джеймса Линда оказались настолько безнадежно забыты или, что гораздо хуже, высокомерно отвергнуты столетие спустя? Частично это объясняется тем, что врачи на самом деле не понимали, каким именно образом плоды цитрусовых защищают от цинги. Другими словами, медиатор не был им известен.

Суточный рацион участников похода Скотта к Южному полюсу: шоколад, пеммикан (блюдо из сушеного мяса), сахар, галеты, масло, чай. Очевидно полное отсутствие фруктов, содержащих витамин С.

Со времен Линда считалось (хоть это и не было доказано) что плоды цитрусовых предотвращают цингу за счет того, что они кислые. Другими словами, врачи считали, что процесс осуществляется в соответствии со следующей каузальной диаграммой: *плоды цитрусовых → кислотность → цинга*.

С этой точки зрения сгодится любой продукт, содержащий достаточное количество любой кислоты, даже кока-кола (хотя она тогда еще и не была изобретена). Сначала моряки брали с собой испанские лимоны: потом, из соображений экономии, их заменили на вест-индские лаймы, которые, хотя и были не менее кислые, чем испанские лимоны, содержали вчетверо меньше витамина С. Затем все пошло еще хуже: моряки стали «очищать» сок лайма, уваривая его, и тем самым окончательно разрушая весь тот витамин С, который в нем еще оставался. Другими словами, они выводили из строя медиатор.

Когда матросы в арктической экспедиции 1875 года заболели цингой, несмотря на то что принимали сок лайма, врачебное сообщество пришло в замешательство. Было известно, что те, кто ел сырое или свежее мясо, не заболели цингой, а те, кто питался мясными консервами, заболели. Кётлиц и прочие решили, что причиной заболевания было неправильно законсервированное мясо. Сэр Элмрот Райт выдумал теорию, согласно которой бактерии в испортившемся (предположительно) мясе вызывают «птомаиновое отравление», которое в свою очередь приводит к цинге. Одновременно с этим теорию, что от цинги можно спастись плодами цитрусовых, отправили на свалку.

Проблема оставалась нерешенной до тех пор, пока не был открыт истинный медиатор. В 1912 году польский биохимик Казимеж Функ предположил существование особых, необходимых организму в микродозах питательных веществ, которые он назвал витаминами. К 1930 году Альберт Сент-Дьёрдьи сумел выделить то вещество, которое предотвращает цингу.

Оно действительно оказалось кислотой, но кислотой особой, которую мы сегодня называем аскорбиновой (противоцинготной) или, проще, витамином С. В 1937 году Сент-Дьёрдьи получил за свое открытие Нобелевскую премию. Благодаря ему сегодня нам известен истинный каузальный путь: *плоды цитрусовых → витамин С → цинга*. Я думаю, что вполне резонно предполагать, что этот каузальный путь ученые уже не забудут. И я надеюсь, что читатели согласятся, что анализ опосредования — это не просто абстрактное математическое упражнение.

Наследственность или воспитание: трагедия Барбаты Бёркс

Насколько мне известно, первым человеком, который смог однозначно представить медиатор посредством диаграммы, была студентка — выпускница Стэнфорда Барбара Бёркс в 1926 году. Одна из первых женщин-ученых, незаслуженно забытая, она истинная героиня этой книги. Есть основания полагать, что она изобрела путевые диаграммы независимо от Сьюалла Райта. А в том, что касается опосредования, она ушла дальше него и на десятилетия опередила свое время.

Областью интересов Бёркс на протяжении всей ее, к несчастью, очень короткой научной карьеры было определение относительной важности наследственности и воспитания в формировании интеллектуальных способностей человека. Ее руководителем в Стэнфордском университете был Льюис Термен — психолог, прославившийся разработкой шкалы интеллекта Стэнфорд — Бине и полагавший, что умственные способности наследуются, а не приобретаются. Важно помнить, что те времена были пиком евгеники, в наше время полностью дискредитировавшей себя, но тогда легитимизированной активными исследованиями таких людей, как Фрэнсис Гальтон, Карл Пирсон и Термен.

Спору «наследственное или приобретенное», конечно, очень много лет, и он продолжался много лет и после Бёркс.

Ее уникальный вклад заключался в том, что она смогла свести вопрос к каузальной диаграмме (рис. 55), которую она использовала, чтобы задаться вопросом (и получить на него ответ): какая часть каузального воздействия обусловлена прямым путем *умственные способности родителей → умственные способности ребенка* (наследственное) и какая — непрямым путем *умственные способности родителей → положение в обществе → умственные способности ребенка* (приобретенное)? В этой диаграмме у Бёркс некоторые стрелки направлены в обе стороны, обозначая либо взаимную причинность, либо неопределенность направления причинности. Для простоты мы предположим, что основное воздействие обеих стрелок идет слева направо, что превращает положение в обществе в опосредующую переменную (медиатор). Таким образом, умные родители занимают в обществе более высокое положение, которое, в свою очередь, позволяет их ребенку лучше развить свои умственные способности. Переменная *X* здесь представляет «другие отдаленные причины, не поддающиеся измерению».

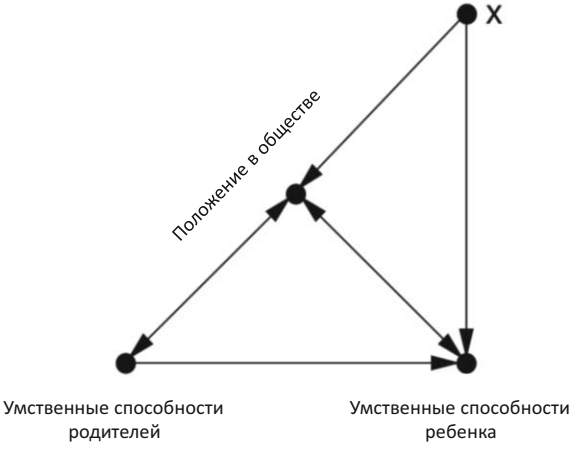


Рис. 55. Спор «наследственное или приобретенное» в интерпретации Барбары Бёркс

В своей диссертации Бёркс проанализировала данные, собранные ею в ходе визитов в 204 приемные семьи, в которых дети предположительно получали от приемных родителей только преимущества воспитания, а не преимущества наследственности. Дети из всех этих семей, а также из контрольной группы в 105 семей без приемных детей, выполняли тест на айкью. В дополнение к тестам заполнялась анкета, в которой различные аспекты условий, в которых росли дети, оценивались по системе баллов. Используя полученные данные и путевой анализ, она подсчитала прямое воздействие родительского ай-кью на интеллект детей и обнаружила, что только 35%, или около $\frac{1}{3}$, изменчивости по ай-кью определяется наследственностью. Другими словами, у родителей с ай-кью на 15 пунктов выше среднего родные дети в среднем оказываются с ай-кью только на 5 пунктов выше среднего.

Научные интересы Барбары Бёркс заключались в разделении наследственной и средовой компонент изменчивости по умственным способностям. Она была первым исследователем после Сьюалла Райта, использовавшим путевые диаграммы, и в некоторых аспектах предвосхитила его.

Как ученица Термена, Бёркс, должно быть, была разочарована, увидев, что воздействие наследственности так мало (надо отметить, что ее оценки вполне выдержали проверку временем). Поэтому она задалась вопросом правомерности применения принятого тогда метода анализа, согласно которому вводились поправки по положению в обществе: «Истинная мера вклада причины в следствие искажается, — пишет она, — если мы сделали постоянными переменные, которые отчасти или полностью *являются следствием любого из двух факторов, отношения между которыми мы измеряем, или третьих отдаленных неизмеренных причин, которые также действуют на любой из двух этих изолированных факторов* (выделено в оригинале)». Другими словами, если нас интересует полное воздействие умственных способностей родителей на умственные способности ребенка, нам не следует вводить

поправки, делая постоянными те переменные, которые стоят на пути между ними. Но Бёркс не останавливается на этом. Ее критерий, выделенный курсивом, в переводе на современный язык звучит так: искажения оценки возникают, если мы вводим поправки по переменным, которые являются следствиями либо: а) умственных способностей родителей или ребенка; либо: б) неизмеренных причин, влияющих на умственные способности родителей или ребенка (как *X* на рис. 55).

Эти критерии намного опережали свое время и совершенно не похожи на что-либо, о чем писал Сьюалл Райт. На самом деле критерий *б* — один из самых ранних примеров ошибки схождения. Если мы посмотрим на рис. 55, то увидим, что положение в обществе — это переменная схождения (*умственные способности родителей → положение в обществе ← X*). Поэтому введение поправок по положению в обществе открывает путь черного хода *умственные способности родителей → положение в обществе ← X → умственные способности ребенка*. В результате любая оценка прямых и непрямых воздействий будет искажена. Поскольку у статистиков как до, так и после Бёркс не было привычки мыслить при помощи стрелок и диаграмм, они полностью погрязли в мифе о том, что, если обычная корреляция не означает каузации, контролируемая корреляция (или частичные коэффициенты регрессии) — это шаг в направлении каузального объяснения.

Бёркс не была первооткрывательницей эффекта схождения, но можно утверждать, что она первой охарактеризовала его графически. Ее критерий *б* прекрасно применим к примерам *M*-искажения в главе 4. Она первой предупредила об опасностях введения поправок по фактору до воздействия — привычки, которая весь XX век всем деятелям в области статистики казалась безопасной (а некоторые, что удивительно, считают ее безопасной до сих пор).

А теперь попробуйте влезть в шкуру Барбары Бёркс. Вы только что открыли, что все ваши коллеги вводили поправки не по тем переменным. Против вас играют два момента: вы женщина и вы пока еще только студентка. Что делать? Покорно склонить голову, сделать вид, что вы согласны с вековой мудро-

стью и общаться с коллегами на их языке, как бы неадекватен он ни был? Кто угодно поступил бы так, только не Барбара! Свою первую работу она озаглавила «О неадекватности техники частичных и множественных корреляций» и начала ее, заявив: «Логические рассуждения приводят к заключению, что техники частичных и множественных корреляций чреваты опасностями, которые серьезно сокращают область их возможного применения». Полемика, навязанная кем-то, у кого еще даже нет научной степени! Как писал Термен, «ее способностям в каком-то смысле противостояла ее тенденция идти против течения. Я полагаю, что проблема отчасти заключалась в том, что она отстаивала свои идеи несколько более агрессивно, чем многие преподаватели и студенты-мужчины готовы были бы терпеть». Очевидно, она обогнала свое время по многим направлениям одновременно.

Вполне вероятно, что Бёркс изобрела путевые диаграммы независимо от Сьюалла Райта, который опередил ее всего на шесть лет. Мы с уверенностью можем утверждать, что ни в каком университетском курсе их быть не могло. Рис. 55 — первое появление путевой диаграммы вне работы Райта и первое в истории, затрагивающее область общественных наук и наук о поведении. Верно, что она упоминает Райта в самом конце своей статьи от 1926 года, но выглядит это упоминание так, будто она добавила его в последний момент. Я подозреваю, что она обнаружила диаграммы Райта только после того, как нарисовала свою собственную, скорее всего, по указанию Термена или строгого рецензента.

Можно только фантазировать, кем бы смогла стать Бёркс, не окажись она жертвой своего времени. После получения докторской степени ей так и не удалось получить должность профессора университета, для которой ее квалификация была несомненно достаточной. Ей пришлось довольствоваться более шаткими позициями, например, в Институте Карнеги. В 1942 году она вышла замуж, и логично было бы ожидать, что это изменит ее судьбу к лучшему, но на самом деле она впала в глубокую депрессию. «Я убеждена, что (уж не знаю, права она была в этом или нет) она была уверена: в ее психике

происходит нечто фатальное, с чем она не сможет справиться, — написала Термену ее мать, Фрэнсис Бёркс. — Поэтому из любви и сострадания ко всем нам она решила избавить нас от мучительного зрелища своего трагического распада». 25 мая 1943 года, в 40 лет, она лишила себя жизни, спрыгнув с моста Джорджа Вашингтона в Нью-Йорке.

Но идеи умеют выживать и после трагедий. Когда социологи Хьюберт Блэлок и Отис Дункан возродили путевой анализ в 60-х годах XX века, работа Бёркс послужила им источником вдохновения. Дункан объяснил, что один из его научных руководителей Уильям Филдинг Огборн коротко упомянул путевые коэффициенты в 1946 году в своей лекции по частичным корреляциям. «Огборн цитировал краткое сообщение Райта, то, которое посвящено материалам Бёркс, и я добыл этот репринт», — сообщает Дункан.

Вот как оно вышло — работа Бёркс от 1926 года вызвала интерес Райта к неправомерному использованию частичных корреляций. Ответ Райта через 20 лет попал в лекцию Огборна и запомнился Дункану. Еще через 20 лет, когда Дункан прочитал работу Блэлока по путевым диаграммам, в его памяти всплыло это полузабытое воспоминание из студенческих лет. Просто невероятно и восхитительно, как идея хрупкой бабочкой пролетела почти незамеченной через два поколения, чтобы в итоге триумфально выйти на свет!

В поисках языка (парадокс абитуриентов Беркли)

Несмотря на раннюю работу Бёркс, спустя полстолетия статистикам все еще было сложно даже выразить саму идею прямых и непрямых воздействий, не говоря уже о том, чтобы оценить их. Пример, которым я проиллюстрирую это, похож на парадокс Симпсона, но с хитрым вывертом.

В 1973 году Юджин Хаммель, заместитель декана в Калифорнийском университете, обнаружил тревожную тенденцию в соотношении мужчин и женщин среди поступающих в вуз.

Согласно его данным, из подающих документы в высшую школу Беркли мужчин зачислили 44%, в то время как из числа абитуриентов-женщин зачислили только 35%. В то время дискриминация по полу начала привлекать широкое общественное внимание, и Хаммель не хотел сидеть сложа руки до тех пор, пока у кого-нибудь не появятся по этому поводу вопросы. Он решил исследовать причины такой диспропорции.

Решения по зачислению абитуриентов в Беркли, как и в других университетах, принимаются отдельными факультетами, а не университетом в целом. Поэтому показалось разумным рассмотреть данные по поступлениям по отдельным факультетам. Однако, сделав это, Хаммель обнаружил удивительный факт. Оказалось, что на каждом отдельном факультете решения приемной комиссии были в пользу женщин, а не мужчин. Как это вообще возможно?

На этом этапе Хаммель поступил разумно: он пригласил профессионала в статистике. Питер Бикель, которого попросили посмотреть на эти данные, немедленно понял, что перед ним разновидность парадокса Симпсона. Как мы видели в главе 6, это тенденция, которая направлена в одну сторону в каждой отдельной группе данной популяции (в приведенном случае каждый факультет предпочитает зачислять женщин), но в противоположную сторону, если рассматривать всю популяцию в целом (в целом по университету видимые предпочтения оказываются мужчинам). В главе 6 мы также видели, что правильное решение парадокса очень зависит от вопроса, которым мы задаемся. В этом случае вопрос ясен: виновен ли университет (или кто-то из сотрудников университета) в сознательной дискриминации женщин?

Когда я впервые рассказал своей жене об этом случае, первой ее реакцией было: «Этого не может быть. Если каждый отдельный факультет дискриминирует мужчин, университет в целом не может дискриминировать женщин». И она была права! Этот парадокс задевает наше понимание дискриминации, которое представляет собой каузальную концепцию, включающую преференции в связи с заявленным полом абитуриента. Если все участники предпочитают один пол другому, вся группа в целом

должна демонстрировать те же предпочтения. Если кажется, что данные свидетельствуют об обратном, это означает, что мы обрабатываем данные неправильно, а не в соответствии с логикой причинности. Только с такой логикой и с внятной каузальной схемой мы определим, виновен университет или нет.

На самом деле Бикель и Хаммель сумели найти каузальную схему, которая их полностью удовлетворила. Они написали об этом статью, опубликованную в журнале «Сайенс» в 1975 году, предположив простое объяснение: абитуриентов-женщин отвергали чаще просто потому, что они подавали заявления на факультеты, на которые вообще сложнее поступить.

Если подробнее, то на факультеты гуманитарных и общественных наук пытаются поступить в основном абитуриенты женского пола. На этих факультетах они сталкиваются с двойной неприятностью: число подающих документы на них больше, а число мест — меньше. В свою очередь, женщины намного реже подавали документы на факультеты вроде инженерной механики, куда поступить было легче. У этих факультетов было больше денег и больше мест для студентов, короче говоря, больший процент подавших документы оказывался принят.

Почему женщины подавали документы на те факультеты, на которые сложнее поступить? Возможно, от поступления на технические факультеты их отпугивало то, что на них требуется лучше знать математику, или то, что они традиционно воспринимаются как «мужские». Вероятно, это результат дискриминации на более ранних стадиях обучения: общество имело тенденцию отталкивать женщин от технических дисциплин, пример Барбары Бёркс недвусмысленно это демонстрирует. Но указанные обстоятельства университет Беркли не мог изменить, поэтому дискриминации со стороны вуза не было. Бикель и Хаммель подвели итог: «Университет в целом не причастен к дискриминации против абитуриентов-женщин».

Я хочу хотя бы вкратце отметить точность языка, использованного Бикелем в этой статье. Он проводит тщательные различия между двумя терминами, которые в бытовом английском часто кажутся синонимами: *bias* («перекос») и *discrimination*

(«дискриминация»). Он определяет перекос как «характер ассоциации между конкретным решением и полом конкретного абитуриента». Обратите внимание на слова «характер ассоциации». Они говорят нам о том, что перекос — явление первой ступени Лестницы Причинности. В свою очередь, дискриминацию он характеризует как «принятие решения в зависимости от пола абитуриента тогда, когда это не имеет отношения к квалификационным требованиям для зачисления». Формулировки «принятие решения», «в зависимости от», «не имеет отношения» просто благоухают причинностью, несмотря на то что Бикель в 1975 году еще не мог решиться произнести это слово вслух. Дискриминация, в отличие от перекоса, находится на второй или третьей ступени Лестницы Причинности.

Анализируя данные, Бикель почувствовал, что они должны быть разделены на страты по факультетам, потому что именно отдельные факультеты были тем уровнем, на котором принимались решения. Было ли это правильно? Чтобы ответить на этот вопрос, начнем с того, что нарисуем каузальную диаграмму (рис. 56). Весьма полезно также посмотреть, как определяет дискриминацию американское прецедентное право. Оно использует контрфактивную терминологию, и это явный признак того, что мы поднялись на третью ступень Лестницы Причинности. В деле Карсона против Вифлеемской сталелитейной корпорации (1996) апелляционный суд седьмого округа писал: «Центральный вопрос в любом деле о дискриминации при приеме на работу заключается в том, предпринял бы наниматель такое же действие, если бы кандидат был человеком другой расы (возраста, пола, вероисповедания, этнического происхождения и т.д.), а все остальное было бы таким же». Это определение четко выражает идею, что мы должны заблокировать, или «заморозить», все каузальные пути, которые ведут от пола к зачислению через любые другие переменные (например, квалификация, предпочитаемые факультеты и т.д.). Другими словами, дискриминация равна прямому воздействию пола на исход решения о зачислении.

Ранее мы видели, что введение поправок по медиатору неправомерно, если мы хотим оценить общее воздействие одной

переменной на другую. Но в случае дискриминации, согласно решению суда, имеет значение не общее, а прямое воздействие. Таким образом, решение Бикеля и Хаммеля оправдано: при допущениях, показанных на рис. 56 они совершенно верно разделили данные по отдельным факультетам, и их результаты представляют собой обоснованную оценку прямого воздействия пола на зачисление. Им удалось успешно справиться с задачей, несмотря на то что терминология прямых и непрямых воздействий не была доступна Бикелю в 1973 году.

Рис. 56. Каузальная диаграмма для парадокса с зачислением в Калифорнийский университет в Беркли (простая версия)

Однако самое интересное в этой истории — не исходная статья, написанная Бикелем и Хаммелем, а дискуссия, которая последовала за ней. После публикации статьи Уильям Крускал из Университета Чикаго написал Бикелю письмо, утверждая, что их объяснение на самом деле не снимает обвинений с Беркли. На самом деле Крускал задался вопросом, насколько вообще любое исследование, основанное исключительно на наблюдениях (в противоположность РКИ) на это способно.

Для меня их переписка — просто невероятное явление. Не так часто нам удается увидеть, как два великих ума сражаются с концепцией (в данном случае с причинностью), для описания которой у них нет адекватного словаря. Позднее Бикель продолжит тему и выиграет «грант гения» от Фонда Макартуров в 1984 году. Но в 1975 году он был еще только в начале своей карьеры, и для него было одновременно честью и вызовом схлестнуться с Крускалом, настоящим великаном в американском статистическом сообществе.

В своем письме к Бикелю Крускал указал на то, что в отношениях между факультетом и результатом зачисления может быть неизмеримый осложнитель, такой как место проживания. Он предоставил численный пример, представляющий гипотетический университет с двумя факультетами, дискриминирующими по полу, в которых данные в результате полностью совпадают с данными в примере Бикеля. Он добился этого, предположив, что оба факультета принимают всех мужчин, живущих в том же штате, и всех женщин из других штатов, и отвергают всех мужчин из других штатов и женщин из того же штата, и это у них единственный критерий для принятия решения о зачислении. Разумеется, такая политика зачисления студентов будет грубым, азбучным примером дискриминации. Но поскольку общие числа принятых и отвергнутых абитуриентов каждого пола были точно такими же, как в примере Бикеля, то по логике последнего пришлось бы признать, что дискриминации нет. Согласно Крускалу, факультеты только кажутся невиновными, потому что Бикель вводил поправки лишь по одной переменной вместо двух.

Крускал указал пальцем на самое слабое место в статье Бикеля: отсутствие четко оправданного критерия, по каким переменным вводить поправки. Крускал не предложил никакого решения, и на самом деле в его письме чувствуется отчаяние от неверия, что оно в принципе реально.

В отличие от Крускала, мы можем нарисовать диаграмму и ясно увидеть, в чем заключается проблема. На рис. 57 представлена диаграмма, соответствующая контрпримеру Крускала. Не кажется ли она вам чем-то знакомой? Так и должно быть! Это совершенно та же диаграмма, что и нарисованная в 1926 году Барбарой Бёркс, только переменные в ней другие.

Тут просится на язык американская пословица «Великие умы мыслят схоже», но, возможно, вернее было бы сказать, что великие проблемы привлекают внимание великих умов.

Крускал утверждал, что при анализе в этой ситуации нужно вводить поправки как по переменной *факультет*, так и по *штату проживания*, и взгляд на рис. 57 объясняет, почему это так. Чтобы заблокировать все пути, кроме прямого,

мы должны стратифицировать данные по факультетам. Таким образом мы закроем непрямые пути *пол → факультет → результаты зачисления*. Но, сделав это, мы открываем побочный путь *пол → факультет ← штат проживания → результаты зачисления* из-за переменной схождения *факультет*. Если мы также введем поправки по переменной *штат проживания*, то закроем этот путь, и поэтому все оставшиеся корреляции должны быть обусловлены прямым (дискриминационным) путем *пол → результаты зачисления*. За отсутствием диаграмм Крускалу пришлось убеждать Бикеля числами, и на самом деле его числа показали то же самое. Если мы вообще не вводим никаких поправок, то процент зачисляемых женщин ниже. Если мы вводим поправки по факультету, то у женщин процент зачисления кажется выше. Если мы вводим поправки и по факультету, и по штату проживания, числа снова покажут более низкий процент зачисления для женщин.

Рис. 57. Каузальная диаграмцма для парадокса зачислений в Калифорнийский университет в Беркли (версия Крускала)

Подобные аргументы показывают нам, почему концепция опосредования вызывала ранее (и до сих пор вызывает) такие подозрения. Она выглядит нестабильной и неуловимой. Результаты зачисления оказываются настроены против женщин, потом против мужчин, потом снова против женщин. В своем ответе Крускалу Бикель продолжал настаивать, что поправка по месту, где принимаются решения (факультет) чем-то отличается от поправки по критерию этого решения (штат проживания). Однако 100%-ной уверенности, судя по всему, у него по этому поводу не было. Он спрашивает несколько беспомощно: «Здесь

я вижу нестатистический вопрос: что такое перекос?» Почему знак перекоса меняется в зависимости от того, как мы его измеряем? На самом деле его идея принципиального различия между перекосом и дискриминацией была верна. Перекос, искажение — это скользкое статистическое понятие, которое может исчезнуть, если нарезать данные не вдоль, а поперек. Дискриминация, как причинностная концепция, отражает реальность, и поэтому должна оставаться неизменной вне зависимости от способа обработки данных.

Фраза, которой не хватало в словаре у них обоих,— «оставаться постоянной». Чтобы заблокировать непрямой путь от пола к результатам зачисления, мы должны зафиксировать значение переменной *факультет* и «покрутить» переменную *пол*. Когда переменная *факультет* принимает фиксированное значение, мы (фигурально выражаясь) не даем абитуриентам выбирать, на какой факультет подавать документы. Поскольку в статистике нет слова для этого понятия, обычно делается нечто с виду похожее: вводится поправка по переменной *факультет*. Именно это сделал Бикель: он стратифицировал данные по факультетам и заключил, что дискриминации нет. Эта процедура законна, когда переменные *факультет* и *результаты зачисления* не имеют осложнителей; в этом случае «видеть» и «делать» — одно и то же. Но Крускал совершенно корректно спросил: «А что, если там есть осложнитель, например штат проживания?» Он, вероятно, и не осознавал, что шел по следам Бёркс, нарисовавшей точно такую же диаграмму.

Невозможно даже передать, насколько часто эта ошибка появлялась в рассуждениях долгие годы — по опосредующей переменной вводились поправки, вместо того чтобы придать ей постоянное значение. По этой причине я называю ее Заблуждением Опосредования. Безусловно, это заблуждение безобидно, если ни опосредующая, ни итоговая переменные не осложнены. Но если осложнители присутствуют, результаты анализа могут привести к совершенно противоположным выводам, как показал численный пример Крускала. Он демонстрирует, как легко исследователь может прийти к выводу об отсутствии дискриминации там, где она есть.

Бёркс и Крускал были исключениями своего времени, признавая Заблуждение Опосредования ошибкой, хотя они и не предлагали для него верного решения. Р. Э. Фишер пал жертвой этого же заблуждения в 1936 году, и сейчас, 80 лет спустя, статистика все еще не справилась с этой проблемой. К счастью, со времен Фишера наблюдается значительный прогресс. Так, эпидемиологи знают, что необходимо следить за осложнителями на пути между опосредующей переменной и итоговой. Однако есть и те, кто отказывается от языка диаграмм (как некоторые экономисты до сих пор), они жалуются и признаются, что для них объяснить, что означает это предупреждение, — просто пытка.

Хорошо, что проблема, которую Крускал однажды назвал «возможно, принципиально неразрешимой», была решена два десятилетия назад. У меня есть странное чувство, что Крускалу это решение понравилось бы, и в своем воображении я демонстрирую ему мощь *do*-исчисления и контрфактивной алгоритмизации. К сожалению, он вышел на пенсию в 1990 году, как раз тогда, когда правила *do*-исчисления постепенно формировались, и умер в 2005 году.

Я уверен, что некоторым читателям интересно, чем в итоге закончилось дело с университетом Беркли. Отвечу: ничем. Хаммель и Бикель были убеждены, что университету не о чем беспокоиться, и, действительно, никаких судебных процессов или федеральных расследований в итоге не проводилось. Данные намекали на обратную дискриминацию, в пользу женщин, и на самом деле этому были явные свидетельства: «В большинстве случаев, в которых женщинам оказывались предпочтения, дело, похоже, обстояло таким образом, что приемные комиссии старались преодолеть многолетний недостаток женщин в их областях науки», — писал Бикель. Всего спустя три года дело об антидискриминационных мерах против другого кампуса Калифорнийского университета дошло до Верховного суда. Если бы Верховный суд отменил антидискриминационные меры, подобные «льготы для женщин» стали бы противозаконными. Однако Верховный суд поддержал их, и случай с Беркли стал исторической вехой.

Истинный мудрец оставит последнее слово не за Верховным судом, а за своей женой. Почему у моей супруги было такое сильное интуитивное убеждение, что университет в целом не может дискриминировать кого-либо, если все его факультеты действуют честно? Это теорема каузального исчисления, похожая на принцип «само собой разумеется». Этот принцип в том виде, в каком его постулировал Леонард Джимми Сэвидж, относится к общему воздействию, в то время как данная теорема верна для прямого воздействия. Само определение прямого воздействия опирается на суммирование прямых воздействий в субпопуляциях.

Коротко и ясно — честность на каждом отдельном месте составляет общую честность. Моя жена была права.

Дэйзи, котята и непрямые воздействия

До сих пор мы обсуждали понятия прямого и непрямого воздействия на смутном и интуитивном уровне, но их точного научного значения я пока не давал. Давно пора устранить это упущение. Начнем с прямого воздействия, потому что это однозначно проще, и мы сможем определить одну из его разновидностей с помощью *do*-исчисления (т.е. на уровне второй ступени Лестницы Причинности). Сначала мы рассмотрим простейший случай, включающий три переменные: экспериментальное воздействие X, результат Y и медиатор M. Мы получаем прямое воздействие X на Y, когда мы «крутим» X, при этом оставляя M постоянным. В контексте примера с парадоксом зачислений в Беркли, мы заставляем всех поступать на исторический факультет, или, другими словами, $do\,(M) = 0$. Мы просим абитуриентов в случайном порядке указывать в анкете свой пол как мужской ($do\,(X) = 1$) или женский ($do\,(X) = 0$) вне зависимости от их настоящего пола. Затем мы наблюдаем разницу в проценте зачисленных в обеих группах. Полученный нами результат называется контролируемым прямым воздействием или КПВ (0). При записи символами:

$$\text{КПВ}\,(0) = P\,(Y = 1 \mid do\,(X = 1),\ do\,(M = 0)) - P\,(Y = 1 \mid do\,(X = 0), do\,(M = 0)). \quad (9.1)$$

Ноль в КПВ (0) означает, что мы сделали так, чтобы опосредующая переменная M приняла нулевое значение. Мы могли бы проделать тот же эксперимент, но заставив всех подать документы на факультет инженерии: $do\,(M = 1)$. Мы обозначим полученное в результате контролируемое прямое воздействие как КПВ (1).

Еще на этом этапе мы видим одно различие между прямым и суммарным воздействиями: у нас есть две версии контролируемого прямого воздействия, КПВ (1) и КПВ (0), какая из них верная? Одно решение — просто указать в отчете обе версии. В самом деле, несложно представить себе, что один факультет дискриминирует женщин, а другой — мужчин, и интересно было бы узнать, как поступает какой из них. В конце концов, изначальное намерение Хаммела в этом и состояло.

Тем не менее я не рекомендовал бы проводить этот эксперимент, и вот почему. Представим себе абитуриента по имени Джо, который всю жизнь мечтал заниматься проектированием и которому в этом эксперименте оказалось случайно назначено подавать документы на исторический факультет. Как человек, на своем веку заседавший в приемных комиссиях, скажу, что всем членам комиссии заявка Джо покажется очень странной. Отличные оценки по теории электромагнитных волн и весьма посредственные по Новейшей истории Европы сильно исказят решение комиссии, вне зависимости от того, будет ли в его анкете в графе «пол» подчеркнуто «муж.» или «жен.». В результате пропорции мужчин и женщин, зачисленных в результате этого эксперимента, вряд ли будут хоть как-то отражать политику зачисления по отношению к тем абитуриентам, которые обычно подают документы на исторический факультет.

К счастью, есть альтернатива, избегающая ловушек этого сверхконтролируемого эксперимента. Мы попросим абитуриентов в графе «пол» указывать рандомизированную информацию, но подавать документы на тот факультет, на который им хотелось бы поступить. Назовем это натуральным прямым

воздействием (НПВ), потому что каждый успешный абитуриент оказывается на факультете своей мечты. «Бы» во фразе намекает на то, что формальное определение НПВ должно быть контрфактивным. Для читателей, любящих математику, вот его определение, выраженное формулой:

$$\text{НПВ} = P\,(YM = M_0 = 1 \mid do\,(X = 1)) - \\ - P\,(YM = M_0 = 1 \mid do\,(X = 0)). \qquad (9.2)$$

Интересен здесь первый член формулы, означающий вероятность, что абитуриентка, выбирая факультет, на который она хотела бы поступить($M = M_0$), будет зачислена, если она укажет свой пол как мужской ($do\,(X = 1)$).

Здесь выбор факультета определяется (отчасти) настоящим полом абитуриента, а решение о зачислении принимается на основе указанного (ложного) пола абитуриента. Поскольку первый нельзя назначить, мы не можем преобразовать этот член таким образом, чтобы он содержал *do*-операторы; нам придется воспользоваться контрафактивным нижним индексом.

Теперь нам известны определения контролируемого прямого воздействия и натурального прямого воздействия, но как нам их сосчитать? Для контролируемого прямого воздействия задача проста; поскольку он выражается в *do*-записи, нам всего лишь потребуется воспользоваться законами *do*-исчисления, чтобы преобразовать *do*-выражения в *see*-выражения (т.е. условные вероятности, которые оцениваются из данных, полученных в результате наблюдений). Натуральное прямое воздействие, однако, представляет собой больший вызов, потому что его нельзя определить в виде *do*-выражения. Для него требуется язык контрфактивных высказываний, и поэтому его невозможно оценить, используя *do*-исчисление. Одним из самых захватывающих прорывов в моей жизни был момент, когда мне удалось очистить формулу для НПВ от всех ее контрфактивных нижних индексов. Результат, названный Формулой Опосредования, делает НПВ действительно очень полезным инструментом, потому что мы в результате оцениваем его только из данных, полученных в результате наблюдений.

Непрямые воздействия, в отличие от прямых воздействий, лишены контролируемых вариантов, потому что нет шанса заблокировать прямой путь, удерживая какую-либо переменную на постоянном уровне. Но у них есть натуральный вариант, натуральное непрямое воздействие (ННВ), определение которого, как и в случае НПВ, использует контрфактивные высказывания. Чтобы сделать определение более наглядным, я использую несколько шутливый пример, который предложил мой соавтор.

Мой соавтор с женой взяли из приюта собаку по имени Дэйзи, неугомонную помесь пуделя и чихуахуа себе на уме. Приучить Дэйзи к соблюдению порядка в доме оказалось не так легко, как ее предшественницу, и спустя несколько недель в доме с ней иногда все еще случались «неожиданности». Однако потом случилось нечто странное. Дана с женой принесли домой на передержку троих котят из приюта, и собачьи «аварии» вдруг прекратились. Котята оставались с ними три недели, и за все это время Дэйзи ни разу не нарушила правила гигиены.

Было ли это просто совпадением, или котята каким-либо образом вдохновили Дэйзи на примерное поведение? Жена Даны предположила, что благодаря котятам Дэйзи почувствовала себя в стае, и поэтому не хотела пачкать там, где у стаи дом. Эта теория снова всплыла на поверхность, когда всего через несколько дней после того, как котята уехали назад в приют, Дэйзи начала оставлять лужи в доме, словно ей ничего никогда не было известно о хороших манерах.

Но затем Дана осознал, что с приездом и отъездом котят менялось и кое-что еще. Пока котята жили в доме, Дэйзи приходилось либо изолировать от них, либо держать под тщательным присмотром. Поэтому ее или надолго запирали в ее собачьем домике, или за ней пристально наблюдали, или даже водили на поводке. При этом обе эти процедуры, и запирание и вождение на поводке, считаются также хорошим способом приучения собак делать свои дела на улице.

Когда котята вернулись в приют, чета Маккензи прекратила непрерывный надзор над собакой и бескультурное поведение вернулось. Дана предположил, что воздействие котят было

непрямым (как в случае теории стаи), а косвенным, опосредованным запиранием в домике и надзором. На рис. 58 показан соответствующий каузальный граф. Начиная с этого момента Дана с женой решили поставить эксперимент. Они вели себя с Дэйзи так, как если бы котята оставались в доме, запирая ее в домике и тщательно присматривая за ней, когда она находилась за его пределами. Если «аварии» прекратятся, можно уверенно заключить, что за это ответственна опосредующая переменная. Если же они не прекратятся, тогда прямое воздействие (стайная психология) окажется более вероятным объяснением.

Конечно, в иерархии научных доказательств их эксперимент был бы сочтен очень шатким — явно не из тех, результаты которых стоит публиковать в научном журнале. Для настоящего эксперимента потребовалось бы изучение нескольких собак, как в присутствии, так и в отсутствие котят. Тем не менее здесь нас интересует каузальная логика эксперимента. Мы пытаемся понять, что случилось бы, если бы котят не было, а опосредующая переменная приняла то значение, какое бы она имела, если бы котята были. Другими словами, мы удаляем котят (интервенция номер один) и строго присматриваем за собакой, будто котята присутствуют (интервенция номер два).

Внимательно присмотревшись к предыдущему абзацу, вы заметите в нем многочисленные «бы» — контрфактивные условия. Котята присутствовали, когда собака изменила свое поведение, но нас интересует, что случилось бы, если бы их не было. Аналогично, если бы котят не было, Дана не присматривал бы за собакой, но нас интересует, что было бы, если бы присматривал.

Рис. 58. Каузальная диаграмма для приучения Дэйзи к порядку

Теперь вы видите, почему статистикам так долго не удавалось дать определение непрямых воздействий. Если даже единственное контрфактивное высказывание было для них чем-то из ряда вон выходящим, то двойные контрфактивы были просто с другой планеты. Однако это определение близко к нашему интуитивному пониманию причинности. Человеческая интуиция работает настолько убедительно, что жена Даны безо всякой специальной подготовки легко поняла логику предполагаемого эксперимента.

Для читателей, не боящихся формул, ниже приводится определение ННВ, которое мы только что дали словами выше:

$$\text{ННВ} = P\,(YM = M_1 = 1 \mid do\,(X = 0)) - \\ - \text{P}\,(YM = M_0 = 1 \mid do\,(X = 0)). \qquad (9.3)$$

Первое P — результат эксперимента с Дэйзи: вероятность того, что приучение к порядку окажется успешным ($Y = 1$), при условии, что мы не заводим других животных в доме ($X = 0$), но придаем опосредующей переменной такое значение, какое она имела бы, если бы они были ($M = M_1$). Мы противопоставляем это вероятности успешного приучения к порядку в «нормальных» условиях, без других животных в доме. Обратите внимание, что контрфактивная величина M_1 должна вычисляться для каждого экспериментального животного отдельно: другим собакам для воспитания могут быть необходимы другие значения переменной *запирание/надзор*. Это выводит непрямое воздействие из области применимости *do*-исчисления. Этот момент также делает эксперимент невыполнимым, потому что экспериментатор не знает M_1 (u) для конкретной собаки u. Тем не менее, приняв, что между M и Y нет осложнителей, натуральное непрямое воздействие все-таки можно подсчитать. Уберем все контрфактивные переменные из ННВ и получим для него Формулу Опосредования, как мы уже делали для НПВ. Это численное выражение, требующее информации с третьей ступени Лестницы Причинности, сокращается тем не менее до выражения, исчисляемого с использованием данных только первой ступени.

Такая редукция возможна только благодаря нашему предположению об отсутствии осложнителей, которое, благодаря определяющему свойству уравнений в структурной каузальной модели, находится на третьей ступени.

Чтобы закончить историю с Дэйзи, сообщу, что результаты эксперимента оказались противоречивы. Неизвестно, следили ли Дана с женой за Дэйзи так же тщательно, как если бы приходилось не пускать ее к котятам (поэтому неясно, было ли переменной M действительно придано значение M_1). Потребовалось терпение и время — несколько месяцев, — и Дэйзи все-таки научилась делать все свои дела на улице.

Но даже и в этом случае в истории с Дэйзи есть несколько важных уроков. Будучи готовым к возможности столкнуться с опосредующей переменной, Дана смог предположить другой каузальный механизм. Из него было выведено важное практическое следствие: ему и жене не пришлось держать в доме дополнительных животных, составляющих стаю, в течение всей жизни Дэйзи.

Опосредование в линейной стране чудес

Когда вы впервые сталкиваетесь с контрфактивными величинами, вам может показаться странным, что для выражения непрямого воздействия требуется такой громоздкий математический аппарат. В самом деле, скажете вы, непрямое воздействие — это всего-навсего то, что остается, если вычесть прямое воздействие. Иначе мы могли бы написать:

$$\text{Суммарное воздействие} = \text{Прямое воздействие} + \\ + \text{Непрямое воздействие.} \qquad (9.4)$$

Если отвечать на это коротко, то схема не работает в моделях, включающих взаимодействия переменных (иногда говорят «модерацию»). Представим, что некое лекарственное средство стимулирует организм выдел ять фермент, который действует как катализатор: он соединяется с этим лекарствен-

ным средством и лечит болезнь. Суммарный эффект этого препарата будет, конечно же, положительным. Однако прямой его эффект равен нулю, потому что, если мы заблокируем медиатор (например, не давая организму выделять фермент), препарат не подействует. Непрямой эффект также равен нулю, потому что, если пациент не будет получать препарат, а начнет принимать искусственно синтезированный фермент, болезнь тоже не пройдет. Сам по себе фермент не излечивает болезнь. Таким образом, уравнение (9.4) не выполняется: суммарное воздействие положительное, но и прямое, и непрямое воздействия равны нулю.

Тем не менее уравнение (9.4) выполняется автоматически в одной ситуации без необходимости ввода контрафактивных переменных. Это случай линейной каузальной модели, вроде той, которую мы рассматривали в главе 8. Как обсуждалось там, линейные модели не допускают взаимонаправленных взаимодействий между переменными, и это может быть как преимуществом, так и недостатком. Преимуществом в том смысле, что анализ опосредования становится намного проще, а недостатком — если мы захотим описать некий каузальный процесс в реальном мире, в котором такие взаимодействия все-таки присутствуют.

Поскольку анализ опосредования намного проще для линейных моделей, посмотрим, как он осуществляется, и с чем вероятны проблемы. Допустим, у нас есть каузальная диаграмма, выглядящая как рис. 59. Поскольку мы работаем с линейной моделью, мы можем представить силу каждого воздействия одним числом. Метки (путевые коэффициенты) показывают, что увеличение переменной *экспериментальное воздействие* на 1 единицу увеличит переменную *медиатор* на 2 единицы. Аналогично увеличение переменной *медиатор* на 1 единицу увеличит переменную *итог* на 3 единицы, а увеличение *экспериментального* воздействия на 1 единицу увеличит итог на 7 единиц. Все это прямые воздействия. Здесь мы подходим к первой причине того, почему линейные модели так просты: прямые воздействия не зависят от уровня опосредующей пе-

ременной, т.е. КПВ (m) одно и то же для всех значений m и мы можем говорить о единственном прямом воздействии.

Каково же будет суммарное воздействие интервенции, благодаря которой экспериментальное воздействие увеличится на 1 единицу? Во-первых, эта интервенция напрямую вынуждает итог увеличиться на 7 единиц (если мы удерживаем медиатор на постоянном уровне). Она также увеличивает медиатор на 2 единицы. Наконец, поскольку каждое увеличение медиатора на 1 единицу напрямую вызывает увеличение итога на 3 единицы, увеличение медиатора на 2 единицы приведет к дополнительному увеличению итога на 6 единиц. Поэтому суммарное увеличение итога по обоим каузальным путям будет составлять 13 единиц. Первые 7 единиц соответствуют прямому воздействию, а оставшиеся 6 — непрямому воздействию. Проще пареной репы!

Рис. 59. Пример линейной модели (путевая диаграмма) с опосредующей переменной

Итак, если имеется более одного непрямого пути от X к Y, мы оцениваем непрямое воздействие по каждому пути как произведение всех путевых коэффициентов вдоль этого пути. Затем мы получаем суммарное непрямое воздействие, суммируя все непрямые каузальные пути. В итоге суммарное воздействие X на Y равняется сумме прямых и непрямых воздействий. Это правило суммы произведений используется с тех пор, как Сьюалл Райт изобрел путевой анализ, и, строго говоря, оно действительно следует из определения суммарного взаимодействия в терминах *do*-оператора.

В 1986 году Рубен Барон и Дэвид Кенни сформулировали набор принципов для обнаружения и оценки опосредования

в системе уравнений. Основные принципы заключаются, во-первых, в том, что все переменные связаны линейными уравнениями, которые оцениваются путем подбора их в соответствии с данными. Во-вторых, прямые и непрямые воздействия исчисляются путем подбора двух уравнений, соответствующих данным: одного с опосредующей переменной и другого без нее. Значительное изменение коэффициентов в случае, когда вводится опосредующая переменная, считается доказательством наличия опосредования.

Простота и убедительность метода Барона — Кенни снискала ему заслуженные лавры в среде общественных наук. В 2014 году их статья занимала 33-е место сверху в списке самых цитируемых работ за всю историю. Их цитировали чаще, чем Альберта Эйнштейна, чаще, чем Зигмунда Фрейда, чаще почти любого другого ученого, которого только можно вспомнить. Их статья стоит на втором месте среди всех публикаций по психологии и психиатрии, хотя она совсем не о психологии. Она о некаузальном опосредовании.

Беспрецедентная популярность подхода Барона — Кенни, без сомнения, определяется двумя факторами. Во-первых, опосредование — очень востребованная концепция. Наше желание понять, «как действует природа» (т.е. найти M в $X \rightarrow M \rightarrow Y$), вероятно, даже сильнее, чем желание подсчитать его. Во-вторых, этот метод легко редуцируется до процедуры, по простоте подобной кулинарному рецепту, основанной на знакомых концепциях статистики, дисциплины, которая долгое время претендовала на исключительное право на объективность и эмпирическую правомерность. Поэтому почти никто не заметил случившийся при этом гигантский рывок вперед — тот факт, что каузальная величина (опосредование) была определена и оценена чисто статистическими методами.

Тем не менее первые трещины в этом возведенном из регрессий оборонительном сооружении начали появляться еще в начале 2000-х, когда практики попытались обобщить правило суммы произведений для нелинейных систем. Это правило включает два допущения: воздействия вдоль разных путей аддитивны, а путевые коэффициенты вдоль одного пути

перемножаются, и оба они приводят к неверным ответам в нелинейных моделях, как мы увидим ниже.

Это заняло немало времени, но в конце концов практикующие анализ опосредования окончательно пришли в себя. В 2001 году мой покойный друг и коллега Род Макдональд писал: «Я полагаю, что лучший способ обсудить вопрос обнаружения или демонстрации модерации или медиации в регрессии — это отложить всю имеющуюся по этому поводу литературу в сторону и начать с нуля». Самые свежие публикации по опосредованию, похоже, последовали совету Макдональда: контрфактивные и графические методы в них используются более часто и последовательно, чем регрессионный подход. А в 2014 году основоположник метода Барона — Кенни Дэвид Кенни опубликовал новый раздел на своем веб-сайте под названием «Каузальный анализ опосредования». Хотя я бы пока поостерегся назвать его обращенным, Кенни явно осознает, что времена меняются и анализ опосредования вступает в новую эпоху.

Теперь давайте рассмотрим простой пример того, как наши ожидания оказываются неверными, стоит нам только выйти за границы Линейной Страны Чудес. Рассмотрим рис. 60, представляющий собой слегка измененный рис. 59, на котором кандидат на рабочее место решает принять предложение тогда и только тогда, когда обещанное жалование превосходит определенную пороговую сумму, в нашем случае 10 единиц. Предложение по зарплате определяется так, как показано на диаграмме: $7 \times$ *образование* $+ 3 \times$ *навык*. Обратите внимание, что функции, определяющие навык и зарплату, все еще предполагаются линейными, но отношение жалованья к итогу нелинейно, потому что у него есть пороговый эффект.

Давайте подсчитаем для этой модели суммарное, прямое и непрямое воздействия, ассоциированные с увеличением образования на 1 единицу. Суммарное воздействие определенно равно 1, поскольку когда образование меняется с 0 на 1, зарплата поднимается с 0 до $(7 \cdot 1) + (3 \cdot 2) = 13$, что больше порогового значения в 10, и таким образом итог меняется с 0 на 1.

Рис. 60. Опосредование, совмещенное с пороговым эффектом

Вспомним, что натуральное непрямое воздействие — это ожидаемое изменение итога, при учете, что мы не меняем образование, но устанавливаем навык на тот уровень, который он бы принял, если мы увеличили бы образование на 1. Легко увидеть, что в этом случае зарплата увеличивается с 0 до $2 \cdot 3 = 6$. Это ниже, чем пороговое значение 10, поэтому податель заявления откажется и НПВ = 0.

Что же насчет прямого воздействия? Как упоминалось выше, вопрос в том, на каком уровне нам следует удерживать опосредующую переменную. Если навык оставить на том уровне, который был до того, как мы изменили образование, тогда зарплата изменится с 0 до 7 и *итог* = 0. Таким образом, КПВ (0) = 0. Однако, если мы придадим навыку то значение, которое он получает после изменения образования (а именно 2), зарплата изменится с 6 до 13. Это меняет итог с 0 на 1, потому что 13 выше порогового значения для подателя заявления, и он согласится на работу. Итак, КПВ (2) = 1.

Следовательно, прямое воздействие равно 0 или 1 в зависимости от постоянного значения, которое мы придаем опосредующей переменной. В отличие от Линейной Страны Чудес выбор значения медиатора играет огромную роль, и у нас возникает дилемма. Если мы желаем сохранить аддитивный принцип, *суммарное воздействие = прямое воздействие + непрямое воздействие*, нам придется использовать КПВ (2) в качестве определения каузального воздействия. Но это выглядит слишком произвольно и в чем-то даже ненатурально. Если мы предполагаем изменить переменную *образование* и хотим узнать ее прямое воздействие, мы, скорее всего, оставим переменную навык на том уровне, который у нее был. Другими

словами, интуитивно кажется более оправданным использовать в качестве прямого воздействия КПВ (0). Более того, это согласуется с натуральным прямым воздействием в этом примере. Однако тогда мы теряем аддитивность: суммарное воздействие не равно сумме прямого и непрямого воздействий.

Тем не менее — вопреки ожидаемому — несколько видоизмененная разновидность аддитивности сохраняется, не только здесь, но и вообще. Читателям, которые не испугаются небольших подсчетов, возможно, будет интересно сосчитать ННВ для возврата от $X = 1$ до $X = 0$. В этом случае зарплата падает с 13 до 7 и итог меняется с 1 на 0 (т.е. податель заявления отказывается от предложения). Подсчитанное в обратном направлении ННВ = –1.0 Восхищение вызывает тот факт, что *суммарное воздействие* ($X = 0 \rightarrow X = 1$) = НПВ ($X = 0 \rightarrow X = 1$) — ННВ ($X = 1 \rightarrow X = 0$), или в этом случае $1 = 0 - (-1)$. Вы видите версию аддитивного принципа для натуральных воздействий, только в данном случае это оказывается субстрактивный (вычитательный) принцип! Я был невероятно счастлив, когда из анализа стал вырисовываться такой вариант аддитивности, несмотря на нелинейность уравнений.

Немало чернил ушло на споры о самом «правильном» способе обобщения прямых и непрямых воздействий при переходе от линейных к нелинейным моделям. К сожалению, большая часть статей подходит к проблеме с конца. Вместо того чтобы заново, с нуля решить, что мы имеем в виду под прямым и непрямым воздействиями, они начинают с предположения, что нам всего-то нужно немного подправить определения для линейных моделей. Например, в Линейной Стране Чудес мы видели, что непрямое воздействие подается как произведение двух путевых коэффициентов. Поэтому некоторые исследователи попытались определить непрямое воздействие как произведение двух численных выражений, из которых одно измеряет воздействие X на M, а второе — воздействие M на Y. Этот подход стал известен как метод произведения коэффициентов. Однако мы также видели, что в Линейной Стране Чудес непрямое воздействие задается разницей между суммарным воздействием и прямым воздействием. Поэтому

другая, не менее самоотверженная группа исследователей определяла непрямое воздействие как разницу двух численных показателей, один из которых отражал суммарное воздействие, а другой — прямое воздействие. Этот метод стал называться методом разницы коэффициентов.

Какой же из них верен? Ни тот ни другой! Обе группы исследователей спутали процедуру и смысл. Процедура здесь математическая: смысл каузальный. На самом деле проблема еще глубже: исследователи, занимающиеся регрессионным анализом, никогда не рассматривали смысл непрямого воздействия за рамками пузыря линейных моделей. Единственным значением понятия непрямого воздействия был результат алгебраической процедуры (перемножить путевые коэффициенты). Когда эту процедуру у них отобрали, их стало носить ветром, как лодку без якоря.

Один из читателей моей книги «Причинность» в своем письме ко мне прекрасно описывает это чувство растерянности. Мелани Уолл (Колумбийский университет) в свое время преподавала курс математического моделирования биостатистикам и медикам. Однажды она, как обычно, объясняла студентам, как вычислять непрямое воздействие, перемножая прямые путевые коэффициенты. Некий студент спросил ее, что именно имеется в виду под непрямым воздействием. «Я ответила ему то, что отвечала всегда, что непрямое воздействие — это воздействие, которое изменение в X производит на Y через его взаимоотношения с опосредующей переменной Z», — написала мне Уолл. Однако студент оказался очень настойчивым. Он вспомнил, как преподаватель объяснял прямое воздействие как воздействие, которое остается, если мы будем поддерживать опосредующую переменную на постоянном уровне, и спросил: «Тогда что мы удерживаем на постоянном уровне, когда говорим о непрямом воздействии?»

Уолл не знала, что сказать. «Я не уверена, что у меня сейчас есть хороший ответ на этот вопрос, — ответила она. — Как насчет того, что я выясню, что смогу, и сообщу вам?»

Это было в октябре 2001 года, всего через четыре месяца после того, как я представил статью по каузальному опосредо-

ванию на конференции «Неопределенность в искусственном интеллекте» в Сиэтле. Стоит ли говорить, что мне очень хотелось поразить Мелани своим свежим решением ее задачи, и я написал ей то же, что пишу здесь сейчас для вас: «Непрямое воздействие *X* на *Y* — это изменение в *Y*, которое мы наблюдаем, если удерживаем *X* на постоянном уровне и в то же время изменяем *M* до такого уровня, который *M* приняло бы при изменении *X* на единицу».

Я не уверен, что Мелани впечатлилась моим ответом, но ее любознательный студент заставил меня серьезно задуматься о том, как прогрессирует наука в наши времена. Вот сейчас, думал я, прошло 40 лет с тех пор, как Блалок и Дункан ввели путевой анализ в общественные науки. Десятки учебников и сотни научных публикаций по прямым и непрямым воздействиям выходят из печати каждый год, некоторые из них — с заголовками-оксюморонами вроде «Регрессионный подход к опосредованию». Каждое поколение передает следующему из рук в руки мудрость о том, что непрямое воздействие — это всего лишь произведение двух других воздействий, или же разница между суммарным и прямым воздействиями. Никто не осмеливается задать простой вопрос: но что это самое непрямое воздействие означает? Чтобы задать его, вдребезги разбив нашу веру в пророческую роль научного консенсуса, понадобился невинный студент с беззастенчивым нахальством мальчишки из андерсеновского «Нового платья короля».

Знакомьтесь с «если бы»

Далее я расскажу вам историю своего собственного обращения, потому что достаточно долго мне не давал покоя тот же самый вопрос, который озадачил студента Мелани Уолл.

В главе 4 я писал о Джейми Робинсе, пионере статистики и эпидемиологии в Гарвардском университете, который вместе с Сандером Гренландом из Калифорнийского университета в Лос-Анджелесе ввел широкое употребление графических моделей в современной эпидемиологии. Мы сотрудничали

несколько лет, с 1993 по 1995 год, и он навел меня на мысли о проблеме последовательных схем интервенции, что было одним из его главных научных интересов.

За много лет до этого Робинса, как эксперта по производственному здравоохранению и безопасности, попросили выступить в суде с мнением о вероятности того, что воздействие химических веществ на рабочем месте привело к смерти рабочего. Робинс был обескуражен, обнаружив, что в статистике и эпидемиологии совершенно нет инструментов для ответов на подобные вопросы. Это по-прежнему была эра почти полного табу на причинность в статистике. Говорить о ней дозволялось только в случае рандомизированных контролируемых исследований, а по этическим соображениям провести подобный эксперимент на людях, чтобы выяснить последствия воздействия формальдегида, было совершенно невозможно.

Обычно рабочие на фабрике подвергаются воздействию токсичных веществ не однократно, а в течение долгого времени. По этой причине Робинс стал интересоваться всеми случаями, когда уровень воздействия химических веществ менялся с течением времени. Иногда такие воздействия бывают и благотворными, например, при СПИДе препараты применяются в течение многих лет, при этом схемы лечения меняются в зависимости от того, как на них реагирует уровень гликопротеина CD4 у пациента. Как вычленить каузальное воздействие курса лечения, если он состоит из множества стадий и промежуточные переменные (которые используются в качестве контроля) зависят от более ранних стадий лечения? Эти вопросы определили направление карьеры Робинса.

После того как Джейми прилетал ко мне в Калифорнию, услышав о «задаче на салфетке» (см. главу 7), он активно заинтересовался перспективами применения графических методов к последовательным схемам лечения, которые были его коньком. Вместе мы выработали последовательный критерий черного хода для оценки каузального воздействия подобных терапевтических курсов. Это сотрудничество научило меня нескольким важным вещам. В частности, оно показало мне, что

анализировать два действия иногда проще, чем одно, потому что каждое действие соответствует стиранию стрелки на графе и таким образом делает его более разреженным.

Наш критерий черного хода работал с длительным курсом лечения, состоящим из произвольно большого числа *do*-операций. Но даже две операции — это уже интересная математика, в том числе контролируемое прямое воздействие, которое состоит из одного действия, которое «играет» значением экспериментального воздействия, в то время как другое действие удерживает опосредующую переменную на постоянном уровне. Что еще важнее, идея определения прямых воздействий в терминах *do*-операций освободила их из тюрьмы линейных моделей и укоренила в каузальном исчислении. Но по-настоящему я заинтересовался опосредованием только позднее, когда обнаружил, что люди по-прежнему совершают ошибки в самых элементарных вещах, например как в упомянутом выше Заблуждении Опосредования. Меня также огорчало то, что основанное на действии определение прямого воздействия не расширялось на непрямое воздействие. Как сказал студент Мелани Уолл, у нас нет переменной или набора переменных, интервенция по которым могла бы заблокировать прямой путь и оставить непрямой путь действующим. По этой причине непрямое воздействие казалось мне плодом воображения, лишенным независимого значения и только напоминающим нам, что суммарное воздействие может отличаться от прямого воздействия. Я даже писал об этом в таких выражениях в первом издании (2000) своей книги «Причинность». Это был один из трех крупнейших просчетов в моей карьере.

Сейчас, глядя в прошлое, я понимаю, что был ослеплен успехом *do*-исчисления, благодаря которому я уверился в том, что единственный способ заблокировать каузальный путь — это взять переменную и придать ей определенное постоянное значение. Это не так: если у меня есть каузальная модель, я могу манипулировать ей по-разному, творчески, решая, какая переменная «слушает» какую, когда и как. В частности, я могу зафиксировать главную переменную на постоянном уровне, чтобы подавить ее прямое воздействие, и гипотетически, но од-

новременно с этим стимулировать главную переменную, чтобы передать ее воздействие через опосредующую переменную. Это позволит мне выставить переменную экспериментального воздействия (т.е. котят) на ноль и выставить опосредующую переменную на тот уровень, который был бы у нее, если бы уровень переменной *котята* был равен единице. Моя модель процесса, порождающего данные, затем сообщит мне, как подсчитать общее воздействие расщепленной интервенции.

Я в долгу перед одним из читателей первого издания Жаком Хагенаарсом (автором книги «Качественные продольные данные») за совет не оставлять надежду на непрямое воздействие. «Многие эксперты в области общественных наук согласны с наблюдаемым на входе и выходе, разногласия как раз в том, каков механизм», — написал он мне. Но я почти два года не мог сдвинуться с места из-за дилеммы, о которой написал в последнем разделе «Как можно заблокировать прямое воздействие?».

Все эти вопросы пришли к неожиданному разрешению, близкому к божественному откровению, когда я прочел юридическое определение дискриминации, которое я цитировал в этой главе ранее: «… если бы нанимаемый был другой расы… а все остальное было бы точно таким же». Вот она — суть проблемы! Это игра «в понарошку». Мы поступаем с каждым индивидуумом по ее или его заслугам, и мы сохраняем все характеристики этого индивидуума на том уровне, на котором они были до изменения в экспериментальной переменной.

Как это разрешает нашу дилемму? Это означает, в первую очередь, что нам придется заново дать определения как прямого, так и непрямого воздействий. Для прямого воздействия мы позволяем опосредующей переменной принять то значение, которое она имела бы — для каждого индивидуума — в отсутствие экспериментального воздействия, и фиксируем ее в этой точке. Теперь мы «играем» экспериментальной переменной и отмечаем разницу. Эта процедура отличается от контролируемого прямого воздействия, описанного ранее, где опосредующая переменная фиксируется на одном и том же уровне для всех. Поскольку мы позволяем опосредующей переменной принимать ее естественные, «натуральные» значения, я на-

зываю это натуральным прямым воздействием. Аналогично, для натурального непрямого воздействия я сначала исключаю действие экспериментальной переменной для всех и каждого, а затем позволяю опосредующей переменной принять для каждого индивидуума то значение, которое она бы приобрела в присутствии экспериментального воздействия. В конце я опять отмечаю наблюдаемые различия.

Я не знаю, помогло бы законодательное определение дискриминации вам либо кому-либо еще пойти тем же путем, что и я. Однако к 2000 году я уже владел контрфактивным языком, как своим родным. Научившись читать контрфактивные высказывания в каузальных моделях, я понял, что это всего лишь количественные данные, которые подсчитываются с помощью невинных операций с уравнениями или диаграммами. Они как таковые оказались готовы к заключению в математическую формулу. Все, что мне понадобилось, — это ухватить «если бы».

В одну секунду я понял, что каждое прямое и непрямое воздействие можно перевести на язык контрфактивных выражений. Как только я понял, как это делается, выведение формулы для оценки натуральных прямых и непрямых воздействий из данных и определения легитимности этой процедуры получилось по щелчку пальцев. Что важно, эта формула не строит предположений о специфике функциональной формы отношений между X, *M* и *Y*. Нам удалось сбежать из Линейной Страны Чудес.

Новое правило я назвал Формулой Опосредования, хотя на самом деле формул две: одна для натурального прямого воздействия, другая для натурального непрямого воздействия. При условии некоторых вполне прозрачных допущений, эксплицитно выраженных в графе, она рассказывает, как их оценить из имеющихся данных. Например, в ситуации, подобной воспроизведенной на рис. 56, где между переменными нет осложнителей, а *M* — опосредующая переменная между экспериментальной переменной X и результатом *Y*:

$$\text{НПВ} = \sum_m (P(M = m \mid P = 1) - P(M = m \mid X = 0)) * \\ * P(Y = 1 \mid X = 0, M = m). \qquad (9.5)$$

Трактовка этой формулы весьма познавательна. Выражение в скобках означает воздействие X на *M*, а следующее за ним выражение — воздействие *M* на *Y* (когда X = 0). Таким образом, она отражает происхождение идеи произведения коэффициентов, выраженной в виде произведения двух нелинейных воздействий. Обратите внимание также на то, что, в отличие от уравнения (9.3), уравнение (9.5) не содержит нижних индексов и *do*-операторов и, следовательно, оценивается из данных первого уровня причинности.

Неважно, кто вы — ученый в лаборатории или ребенок на велосипеде, вас всегда будет радовать тот факт, что сегодня вы научились чему-то, чего не умели вчера. И именно эту радость я ощущал, когда Формула Опосредования впервые появилась на бумаге. Мне теперь с первого взгляда было видно все о прямых и непрямых воздействиях: что нужно, чтобы увеличить или уменьшить их, когда их оценивают из данных, полученных в результате наблюдений или интервенций, и когда мы можем заявить, что опосредующая переменная «виновна» в передаче наблюдаемых изменений к итоговой переменной. Отношения между причиной и следствием бывают линейными или нелинейными, численными или логическими. Ранее каждый из этих случаев приходилось рассматривать отдельно, если, конечно, о них упоминали вообще. Теперь единая формула годится для любого из них.

Если у нас есть верные данные и верная модель, мы способны определить, виновен ли наниматель в дискриминации или какие осложнители удержат нас от этого вывода. По данным Барбары Бёркс мы оценим, какая часть ай-кью ребенка определяется наследственностью, а какая — воспитанием. Мы даже высчитаем процент общего воздействия, *объясняемый* опосредованием, и процент, *определяемый* опосредованием, — две взаимодополняющие концепции, которые в линейных моделях сливаются в одну.

После того как мне удалось записать контрфактивное определение прямых и непрямых воздействий, я узнал, что я не первым пришел к этой идее. Робинс и Гренланд побывали там до меня, еще в 1992 году. Но их статья описывает концепцию

натурального воздействия словами, не сводя их к математической формуле. Что важнее, они отнеслись к идее натуральных воздействий в целом пессимистически и постулировали, что такие воздействия нельзя оценить даже по экспериментальным исследованиям и уж точно не по исследованиям, основанным на наблюдениях. Это утверждение удержало других исследователей от изучения потенциала натуральных воздействий. Сложно сказать, смогли ли бы Робинс и Гренланд перейти к более оптимистичной точке зрения, если бы они пошли чуть дальше и выразили натуральное воздействие в виде формулы на контрфактивном языке. Для меня этот дополнительный шаг оказался решающим.

У них, вероятно, был еще один повод для пессимизма, с которым я не согласен, но попробую обсудить. Они изучили контрфактивное определение натурального воздействия и увидели, что оно сочетает в себе информацию из двух разных миров, одного, в котором вы удерживаете экспериментальную переменную на нуле, и другого, в котором вы меняете опосредующую переменную на то значение, которое она приняла бы, если бы вы выставили экспериментальную переменную на единицу. Поскольку это условие пересечения миров нереально выполнить ни в одном эксперименте, ученые решили, что оно вне игры. В этом разница их и моего философских подходов.

Они полагают, что легитимизировать причинностные связи можно, только воспроизведя рандомизированное исследование наиболее точно, основываясь на предположении, что это единственно вероятный путь к научной истине. Я же верю, что должны быть и иные пути, чья правомерность происходит из сочетания данных и установленных (или предполагаемых) научных знаний. В этой связи доступны методы и более мощные, чем РКИ, основанные на допущениях третьей ступени, и я не боюсь их использовать. Там, где они зажигают красный свет, останавливая исследователей, я зажигаю зеленый — Формулу Опосредования: если вам годятся эти допущения, то смотрите, что можно сделать! К сожалению, красный свет на светофоре Робинса и Гренланда удержал область опосре-

дования от дальнейшего развития в течение долгих девяти полных лет.

Многих людей формулы пугают, им кажется, что они скорее скрывают информацию, чем делают ее доступной. Однако для математика или для того, кто сумел научиться математическому мышлению, верно как раз обратное. Формула объясняет все: она не оставляет сомнений и двусмысленностей. Читая научную статью, я часто ловлю себя на том, что перепрыгиваю от формулы к формуле, пропуская текст. Для меня формула — это хорошо пропеченная идея. Слова — это сырое тесто, которое только ставят в печь.

Формула служит двум целям, одна из них практическая, вторая социальная. С практической точки зрения студенты или коллеги могут пользоваться ей как рецептом. Рецепт может быть простым или сложным, но в итоге он обещает вам, что, если вы будете следовать пошаговой инструкции, вы получите натуральные прямое и непрямое воздействия, конечно, в том случае, если ваша каузальная модель адекватно отражает реальный мир.

Вторая цель более тонкая и сложно вербализуемая. У меня был друг из Израиля, известный художник. Однажды я приехал к нему в студию, чтобы приобрести одну из его картин, и его полотна были везде — сотни под кроватью, десятки на кухне. Стоили они в диапазоне от 300 до 500 долларов, и выбрать из них одну оказалось нелегкой задачей. Наконец я показал на ту, что висела на стене, и сказал: «Мне нравится вот эта». «Эта стоит пять тысяч долларов», — ответил он. «Как так?» — удивился я, недоумевая и даже немного протестуя. Художник ответил: «Эта в раме». Мне потребовалось несколько минут, чтобы понять, что он имел в виду. Эта картина стоила дорого не потому, что ее вставили в раму. Ее вставили в раму потому, что она была ценной. Из сотен работ в студии автор выбрал и вставил в раму именно ее. Она лучше всего выражала то, над чем он работал на других полотнах, и на ней стояла печать законченности — рамка.

Это вторая цель формулы. Это общественный договор. Она вставляет идею в рамку и говорит: «Это что-то, что я считаю важным. Это нечто, чем стоит поделиться».

Вот поэтому я решил вставить в рамку Формулу Опосредования. Ей стоит делиться, потому что для меня и для многих таких, как я, она представляет собой решение 100-летней дилеммы. И она важна, потому что дает практический инструмент для идентификации механизмов и анализа их относительной важности. Это социальный договор, выраженный Формулой опосредования.

Как только утвердилось мнение, что нелинейный анализ опосредования возможен, исследования в этой области стали множиться как грибы. Если вы доберетесь до базы данных по академическим публикациям и предпримете поиск по заголовкам со словами «анализ опосредования», то до 2004 года вы не найдете практически ничего. Затем будет семь статей в год, потом десять, потом двадцать: сейчас же на эту тему публикуется более сотни работ в год. Я хотел бы закончить эту главу тремя примерами, которые, я надеюсь, хорошо проиллюстрируют разнообразие возможностей, которое открывает нам анализ опосредования.

Примеры исследований опосредования

«Алгебра для всех»: образовательная программа и ее побочные эффекты.

Для государственных школ Чикаго были характерны те же, неразрешимые, на первый взгляд, проблемы, что и для большинства систем школьного образования в мегаполисах: высокий уровень бедности, низкий бюджет и значительная разница в успеваемости между студентами разного расового и этнического происхождения. В 1988 году тогдашний министр образования Уильям Беннет назвал школы Чикаго худшими во всей стране.

Однако в 90-е годы XX века при новом руководстве государственные школы Чикаго предприняли ряд реформ и из «худших

в стране» превратились в «ведущие в стране». Некоторые из руководителей этих преобразований обрели всеамериканскую известность, например Арне Дункан, который стал министром образования при президенте Бараке Обаме.

Одним из нововведений, появившихся еще до Дункана, была политика, принятая в 1997 году, отменяющая корректирующие курсы в высшей школе и требующая, чтобы все девятиклассники проходили курсы на уровне подготовки в колледж, такие как «Английский I» и «Алгебра I». Математическая часть этой образовательной программы называлась «Алгебра для всех».

Увенчалась ли «Алгебра для всех» успехом? Оказалось, что на этот вопрос неожиданно сложно ответить. Обнаружились как хорошие, так и плохие новости. Хорошие состояли в том, что результаты экзаменов действительно улучшились. Оценки по математике выросли на 7,8 балла за три года, что представляет собой статистически значимое различие, эквивалентное тому, что примерно 75% студентов получает на экзамене баллы выше того среднего значения, которое наблюдалось до внедрения программы. Однако перед тем, как заводить разговор о причинности, нам нужно исключить осложнители, а в этом случае имелся один весьма серьезный. К 1997 году квалификация школьников, поступающих в девятый класс, уже улучшилась благодаря более ранним изменениям в программе восьмого класса. Таким образом, в этом случае мы не сравниваем яблоки с яблоками. Поскольку эти школьники пришли в девятый класс с уже более глубокими знаниями, чем школьники из 1994 года, их положительные оценки могли объясняться улучшенной программой для восьмого класса, а вовсе не «Алгеброй для всех».

Гуанлей Хон, профессор кафедры развития человека в Чикагском университете, исследовала имеющиеся данные и нашла, что улучшения в экзаменационных баллах становятся незначимыми, если принимать во внимание этот осложнитель. На этом этапе Гуанлей Хон легко могла бы прийти к умозаключению о том, что «Алгебра для всех» успеха не имела, но она этого не сделала, потому что следовало принять во внимание еще

один фактор — на этот раз не осложнитель, а опосредующую переменную.

Любой хороший учитель знает, что успехи школьников зависят не только от того, чему их учат, но и от того, как именно учат. Когда стали вводить программу «Алгебра для всех», изменилась не только программа девятого класса. Хуже успевающие школьники оказались в одних классах с хорошо успевающими, и им оказалось трудно нагонять. Это привело к целому ряду негативных последствий: разочарованиям, прогулам и, конечно, снижению оценок на экзаменах. Кроме этого, в классах, где вместе учились школьники разной успеваемости, отстающие ученики, получали меньше внимания со стороны учителя, чем в классах, усредненных по успеваемости. Наконец, преподаватели, вероятно, внутренне сопротивлялись предъявляемым к ним новым требованиям. Учителям, привыкшим вести курс «Алгебра I», вероятно, не доводилось учить плохо успевающих школьников, а учителя, которые много занимались с отстающими, скорее всего, не так хорошо умели преподавать алгебру. Все перечисленное и составляло непредвиденные побочные эффекты курса «Алгебра для всех». Анализ опосредования идеально подходит для оценки побочных эффектов.

В итоге Хон предположила, что среда обучения в классе изменилась и сильно повлияла на результаты данной интервенции. Другими словами, она постулировала, что ситуация соответствует каузальной диаграмме на рис. 61. Влияние среды (которое Хон измеряла как средний уровень знания предмета всех учеников в классе) действует как опосредующая переменная между интервенцией «Алгебра для всех» и итоговыми результатами учеников. Вопрос, как обычно в случае анализа опосредования, в том, какая часть воздействия образовательной программы прямая, а какая косвенная. Интересно, что два воздействия были противоположно направлены. Хон обнаружила, что прямое воздействие было положительно: новая образовательная программа прямо вела к увеличению итогового экзаменационного балла на 2,7 единиц. Это, по крайней мере, было изменением в правильном направлении, и оно оказалось статистически значимым (что свидетельствует, что подобное

улучшение может произойти само по себе с низкой вероятностью). Тем не менее из-за изменения среды обучения в классе непрямое воздействие практически обнуляло это улучшение, уменьшая баллы за экзамен на 2,3 единицы. Хон пришла к заключению, что особенности реализации программы «Алгебра для всех» значительно подрывают ее эффективность. Если же сохранить программу обучения, но вернуться к дореформенному принципу составления классов, она должна привести к некоторому небольшому улучшению экзаменационных оценок (и, хотелось бы надеяться, знаний школьников).

Рис. 61. Каузальная диаграмма для эксперимента «Алгебра для всех»

По счастливому совпадению именно это и произошло. В 2003 году государственные школы Чикаго (теперь возглавляемые Дунканом) начали новую реформу, называвшуюся «Алгебра в двойном размере». По ее правилам всем школьникам все-таки приходилось учить алгебру, но у тех учеников, чьи отметки оказывались ниже, чем средние по стране, должен был быть не один урок алгебры в день, а два. Этот момент устранил нежелательный побочный эффект предыдущей реформы. Теперь по крайней мере раз в день плохо успевающие ученики оказывались в той среде обучения, которая была близка к дореформенной. Программа «Алгебра в двойном размере» была просто обречена на успех, и она продолжается по сей день.

Я полагаю, что история «Алгебры для всех» — это успех и в случае анализа опосредования, потому что он объяснил как не особо впечатляющие результаты исходной программы, так и улучшившиеся результаты после ее усовершенствования. Хотя причинно-следственная связь обнаружилась слишком

поздно, чтобы влиять на подгонку образовательной программы в реальном времени, она ответила на наш вопрос «Почему?» после того, как появились факты: почему у исходной программы были такие незначительные результаты? почему следующая реформа сработала лучше? Таким образом, она может управлять реформой в будущем.

Я хочу обратить ваше внимание на еще один интересный момент в работе Хон. Она хорошо знала о подходе Барона и Кенни к прямым и непрямым воздействиям, который я назвал Линейной Страной Чудес. В своей статье она на самом деле выполнила анализ дважды: один раз с использованием разновидности Формулы Опосредования, второй раз с помощью «общепринятых процедур» (ее слова) по Барону и Кенни. Методом Барона — Кенни непрямое воздействие выявить не удалось. Причина этого, скорее всего, в том, о чем я писал раньше: линейные методы не чувствительны к взаимодействиям между экспериментальной и опосредующей переменной. Вероятно, сочетание более сложного материала и менее благоприятной среды обучения в классе вызвало рост разочарования и ухудшение успеваемости у учеников. Убедительно ли это? Я считаю, что да. Алгебра — сложный предмет. Возможно, именно из-за ее сложности двойное внимание учителей по программе «Алгебра в двойном размере» оказалось особенно ценным.

Ген курильщика: опосредование и взаимодействие

В главе 5 я рассказывал про научные и политические войны вокруг проблемы курения в 50-х и 60-х годах XX века. Скептики того времени, включая Р. Э. Фишера и Якоба Ерушалми, утверждали, что очевидная связь между курением и раком легких может быть статистическим артефактом, возникающим из-за осложняющей переменной. Ерушалми полагал, что у курильщиков особый тип личности, а Фишер предполагал существование гена, который предрасполагает людей одновременно к курению и к развитию рака легких.

По иронии судьбы в 2008 году исследователи генома человека обнаружили, что Фишер был в некотором роде прав:

действительно, существует ген, функционирующий очень похожим образом. Это открытие состоялось благодаря новой технике анализа генома, называемой «поиск ассоциаций по всему геному» (Genome-*W*ide Association St*U*d*Y*; G*W*AS). Эта техника — прототип современных исследований по методу больших данных, который позволяет исследователям прочесать весь геном статистически, высматривая варианты генов, которые чаще обнаруживаются у людей с определенными заболеваниями, например шизофренией, или диабетом, или раком легких.

В названии метода важно отметить слово «ассоциация». Этот метод не доказывает наличие причинно-следственной связи: он всего лишь находит гены, ассоциированные с данным заболеванием в данной выборке. Это метод, основанный на данных, а не на гипотезе, поэтому для выявления причинно-следственных связей им пользоваться неудобно.

Хотя предыдущие генетические исследования, базирующиеся на проверке гипотезы, не смогли обнаружить свидетельства связи курения или рака легких с определенными генами, все изменилось в одночасье в 2008 году. В этом году исследователи обнаружили ген в 15-й хромосоме, кодирующий рецепторы никотина в клетках легкого. У этого гена есть официальное название — rs16969968, но это сложно даже для экспертов в геномике. Поэтому его стали называть «Большой» или «мистер Большой» из-за его очень сильной ассоциированности с раком легких. «Среди изучающих курение слова „мистер Большой" понятны всем», — говорит Лаура Биерут, эксперт по проблеме курения в Университете Вашингтона в Сент-Луисе. Я же просто буду называть его геном курильщика.

Тут, как мне кажется, нам должен явиться сварливый призрак Р. Э. Фишера, звеня цепями в подвале и требуя отозвать из печати все то, о чем я писал в главе 5. Да, ген курильщика ассоциирован с раком легких. У него есть два варианта: один обычный, другой более редкий. Люди, у которых оказываются две копии редкого варианта (таких примерно 1/9 всех людей), заболевают раком легких на 77% чаще. Ген курильщика оказывается также связан с поведением при курении. Людям

с опасным вариантом требуется больше никотина, чтобы почувствовать насыщение потребности, и им сложнее бросить курить. Однако есть и хорошие новости: эти люди лучше реагируют на никотинзамещающую терапию, чем люди с обычным, «некурящим» вариантом этого гена.

Открытие такого гена не должно никого сбивать с толку относительно на порядок более значимой причины возникновения рака легких — курения. Мы знаем, что оно ассоциировано с более чем десятикратным увеличением риска получить рак легких. Для сравнения: даже двойная доза гена курильщика увеличивает риск рака легких менее, чем вдвое. Это, конечно, серьезно, но несравнимо с той опасностью, которой вы подвергаете себя (безо всякого смысла), если регулярно курите.

Как обычно, полезно визуализировать обсуждаемый вопрос с помощью каузальной диаграммы. Фишер считал, что ген курильщика (тогда еще совершенно гипотетический) является осложнителем по отношению к курению и раку (рис. 62). Но в качестве осложнителя он и близко не объясняет чрезвычайно сильное воздействие курения на риск рака легких. Это, по сути своей, тот самый аргумент, который в 1959 году Джером Корнфилд привел в своей статье, ставшей решающей в споре вокруг гипотезы генетической предрасположенности.

Рис. 62. Каузальная диаграмма для примера с геном курильщика

Мы сможем легко перерисовать эту каузальную диаграмму так, как показано на рис. 63. В этом случае мы видим, что курение как привычка оказывается опосредующей переменной между геном курильщика и раком легких. Это маленькое изменение точки зрения ставит наш научный спор полностью с ног на голову. Вместо того чтобы спрашивать, вызывает ли

курение рак (теперь мы знаем ответ на этот вопрос), мы задаемся вопросом, как работает неблагоприятная версия гена. Заставляет ли она ее обладателей курить чаще и вдыхать глубже? Или же она каким-то образом делает клетки легких более уязвимыми по отношению к раку? Что сильнее, непрямое воздействие или прямое?

Рис. 63. Каузальная диаграмма для примера с геном курильщика после перегруппировки

От ответа зависит решение проблемы. Если основное воздействие прямое, тогда людей с более опасным вариантом гена следует чаще тестировать на рак легких. Однако, если воздействие в основном непрямое, все упирается в курение. В этом случае мы должны рассказывать таким пациентам о том, что они в группе риска и им важно даже не начинать курить. Если они уже курят, нужно вмешиваться более активно, возможно, предлагая никотинзамещающую терапию.

Тайлер Ван дер Виль, эпидемиолог из Гарвардского университета, прочитал первое сообщение о гене курильщика в журнале «Нэйча» и связался с группой исследователей в Гарварде, возглавляемой Дэвидом Кристиани. С 1992 года Кристиани просил своих пациентов с раком легких, а также их друзей и родственников заполнять опросники и сдавать образцы ДНК в помощь программе исследования. К середине 2000-х годов он собрал данные о 1 800 пациентах с раком легких, а также о 1 400 лицах, не больных раком, в качестве контрольной группы. Когда Ван дер Виль позвонил, образцы ДНК еще лежали в холодильнике. Результаты его анализа на первый взгляд обескураживали. Он обнаружил, что риск рака легкого за счет непрямого воздействия увеличивался всего лишь на величи-

ну от 1 до 3%. Люди с более опасным вариантом гена курили в среднем только на одну сигарету в день больше, что было недостаточно для клинической значимости. Тем не менее их организм реагировал на курение иначе. Воздействие гена курильщика на развитие рака легких было большим и значимым, но только для тех, кто курил. Из этого вытекает интересное предположение относительно подачи результатов. В этом случае контролируемое прямое воздействие КПВ (0) будет, в общем-то, равно нулю: если вы не курите, ген вам не вредит. Однако, если мы придадим опосредующей переменной значение, равное одной или двум пачкам в день, что я обозначу как КПВ (1) или КПВ (2), воздействие гена окажется сильным. Натуральное прямое воздействие усредняет эти контролируемые воздействия. НПВ оказывается положительным, и именно в таком виде оно и было представлено в отчете Ван дер Виля.

Этот пример — классический образец взаимодействия переменных из учебника. В итоге анализ Ван дер Виля доказывает про ген курильщика три важных момента. Во-первых, он лишь незначительно увеличивает потребление сигарет. Во-вторых, он не вызывает рак легких каким-нибудь независимым от курения путем. В-третьих, для тех, кто курит, он значительно увеличивает риск рака легких. Все упирается во взаимодействие гена курильщика и поведения его обладателя.

Как всегда в случае любого нового результата, требуются дополнительные исследования. Биерут указывает на одну проблему с анализом Ван дер Виля и Кристиани: в нем был только один параметр, по которому оценивалось поведение курящих — число выкуренных сигарет в день. Тем не менее вполне может оказаться, что люди с опасным вариантом гена вдыхают дым глубже, чтобы получить бóльшую дозу никотина за одну затяжку. В гарвардском исследовании просто не было данных для проверки этой гипотезы.

Даже несмотря на то, что некоторая неопределенность остается, исследование гена курильщика дает нам представление о том, как может быть устроена персонализированная медицина в будущем. Совершенно ясно, что в этом случае важно, как

взаимодействуют генетика и поведение конкретного человека. Мы все еще не знаем, меняет ли ген поведение человека, как предполагает Биерут, или просто взаимодействует с тем поведением, которое возникло бы независимо от него (как следует из анализа Ван дер Виля). Тем не менее, зная генетический статус людей, мы в состоянии обеспечить людей более точной информацией о тех рисках, которые им угрожают. В будущем каузальные модели, способные выявлять взаимодействия между генами и поведением, либо генами и средой обитания, обязательно станут полезными инструментами эпидемиолога.

Жгуты: скрытая обманчивость

В первый же день службы, прибыв в госпиталь в Багдаде, Джон Крэг, военный хирург, столкнулся с новыми реалиями медицины в условиях военных действий. Рассматривая доску с записями состояний больных в этот день, он заметил дежурной медсестре: «Как интересно — в эту смену применялось наложение жгута».

«Ничего особенного, — ответила медсестра, — у нас каждую смену накладывают».

В самые первые минуты на новой службе Крэг обнаружил огромные изменения, происшедшие в практике лечения ранений в войнах в Афганистане и в Ираке. Хотя их многие столетия использовали как на поле боя, так и на операционном столе, отношение к применению жгутов всегда оставалось противоречивым. Жгут, наложенный на слишком долгое время, приводил к потере конечности. Кроме того, жгуты по необходимости часто изготовлялись из того, что имелось под рукой, поэтому неудивительно, что их эффективность лучше всего описывалась фразой «Авось поможет». После Второй мировой войны жгуты стали считать крайним средством и их наложения официально рекомендовали избегать.

Войны в Ираке и Афганистане радикально поменяли такую политику применения. Это объяснялось двумя моментами: большее количество серьезных травм требовало применения жгутов, и к тому же стали доступны жгуты более

удобных конструкций. В 2005 году главный хирург армии США рекомендовал, чтобы каждый солдат был экипирован медицинским жгутом. К 2006 году, как обнаружил Крэг, в госпитали каждый день доставляли солдат со жгутами, наложенными на ноги или руки, — беспрецедентная ситуация в истории медицины.

С 2002 по 2012 год, по оценкам Крэга, жгуты спасли жизни более чем 2 тысячам военных. Солдаты на фронтах это заметили. Как писал хирург армии США Дэвид Веллинг, «боевые подразделения выходят на опасные миссии со жгутами на конечностях наготове, потому что хотят вовремя остановить опасное кровотечение, если взорвется мина или самодельное взрывное устройство».

Если делать выводы из свидетельств очевидцев и популярности жгутов в бою среди солдат, их польза не должна подвергаться сомнению. Тем не менее крупномасштабных исследований по результатам применения этого средства было немного или даже не было совсем. В мирной жизни случаи, требующие использования жгутов, слишком редки, а в боевых условиях хаос войны не позволяет поставить правильно спланированный научный эксперимент. Но Крэг обнаружил возможность документировать последствия их применения. С помощью медсестер он собирал все данные о всех больных, поступающих в госпиталь с наложенными жгутами, и вскоре его уже прозвали «доктор-жгутовик».

Результаты исследования, опубликованные в 2015 году, оказались далеки от того, что ожидал Крэг. По опубликованным данным, выживаемость пациентов, поступивших в госпиталь с наложенными жгутами, была не выше, чем у пациентов с такими же травмами, но без жгутов. Конечно, теоретизировал Крэг, те из них, кому потребовалось накладывать жгут, могли изначально получить более серьезные увечья. Но даже когда он ввел поправку по этому фактору, сравнивая случаи равной тяжести, жгуты все равно не увеличивали вероятность выживания (табл. 13).

Таблица 13. Данные по выживаемости раненых с применением жгутов и без них

Характер травмы	Всего выживших / без применения жгута		Всего выживших / с применением жгута	
	чел.	%	чел.	%
3 Значительная	502/555	90	416/465	89
4 Серьезная	96/111	86	212/248	85
5 Критическая	16/27	59	4/7	57
Итого	614/693	89	632/720	88

Это не та ситуация, с которой мы сталкивались в парадоксе Симпсона. Не имеет значения, объединяем ли мы данные или разделяем их на страты: в каждой категории опасности травм, как и в объединенной выборке, выживание было несколько выше у солдат, которым не накладывали жгуты (разница в процентах выживших была, однако, слишком мала для статистической достоверности).

Что же пошло не так? Один из возможных вариантов ответа, конечно, что жгуты не помогают. Наша вера в них может быть случаем «ошибки выжившего». Когда солдату накладывают жгут и он выживает, врачи и однополчане говорят: «Жгут спас ему жизнь». Но если солдат выжил после ранения, а жгут не накладывали, никто не скажет: «Ненакладывание жгута спасло ему жизнь». Таким образом жгуты могут обретать незаслуженную славу, а отсутствие интервенции остается незамеченным.

Однако в этом исследовании мог быть и еще один источник искажений, на который указал сам Крэг: врачи собирали данные только по тем пациентам, которые прожили достаточно долго для того, чтобы их успели доставить в больницу. Чтобы понять, почему это происходит, нарисуем каузальную диаграмму (рис. 64).

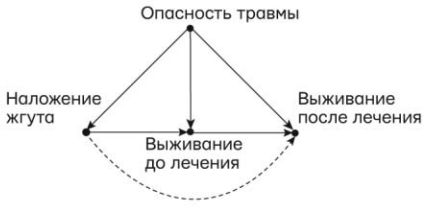


Рис. 64. Каузальная диаграмма для примера со жгутами. Пунктирная линия обозначает гипотетическое причинностное воздействие (не поддерживаемое данными)

На этом рисунке мы видим, что переменная *опасность травмы* является осложнителем для всех трех переменных: экспериментальной (*наложение жгута*), опосредующей (*выживание до лечения*) и итога (*выживание после лечения*). Поэтому оправданно и необходимо вводить поправки по опасности травмы, как Крэг и делал в своей статье.

Однако, поскольку Крэг изучал только пациентов, которые прожили после ранения достаточно долго для того, чтобы их довезли до госпиталя, он таким образом вводил также поправки по опосредующей переменной *выживание до лечения*. На практике он таким образом блокировал непрямой путь от использования жгутов к выживанию после лечения и вычислял прямое воздействие, обозначенное пунктирной стрелкой на рис. 64. Это воздействие практически равно нулю. Тем не менее возможно все же и непрямое воздействие. Если благодаря жгутам больше солдат доживает до госпиталя, тогда наложение жгута — очень желательная интервенция. Это означает, что смысл жгута в том, чтобы доставить пациента до больницы живым: когда это сделано, от него больше нет толку. К сожалению, в имеющихся данных (см. табл. 13) нет ничего, что опровергало или подтверждало бы эту гипотезу.

Уильям Крускал некогда печалился, что нет такого Гомера, который бы воспел подвиги на поле боя статистики. Я бы хотел воспеть научный подвиг Крэга, которой в невообразимо сложных условиях смог мыслить четко, собрать данные и подвергнуть стандартную процедуру научному тестированию. Его

пример, как маяк, освещает путь всем, кто хотел бы заниматься медициной на основе наблюдаемых данных. Особо горькая ирония заключалась в том, что его исследование не могло увенчаться успехом, потому что он никаким образом не мог собрать данные по бойцам, которые погибли, не успев попасть в госпиталь. Было бы замечательно, если бы он смог доказать раз и навсегда, что жгуты спасают людям жизнь. Как писал сам Крэг в электронном письме, «я не сомневаюсь, что наложение — это оправданная мера». Но в итоге ему пришлось доложиться о «нулевом результате», а такие результаты не попадают в газетные заголовки. Тем не менее он заслуживает уважения за здоровые научные инстинкты.

Глава 10
Большие данные, искусственный интеллект и важные вопросы

Все предопределено,
но всегда дается разрешение.
Маймонид (Моше бен Маймон),
1135—1204

Я начал изучать причинно-следственные связи, отправившись по следам аномалии. С помощью байесовских сетей мы научили машины рассуждать, учитывая оттенки серого, и это был важный шаг на пути к мышлению человеческого уровня. Но мы так и не смогли добиться, чтобы машины понимали причины и следствия. Мы не смогли объяснить компьютеру, почему нельзя вызвать дождь, изменив показания барометра. И точно так же не сумели добиться, чтобы он понял, чего ожидать, если один из солдат в расстрельной команде передумает и решит не стрелять. Без способности видеть альтернативные варианты реальности и противопоставлять их существующей реальности машина не может пройти мини-тест Тьюринга. Она не способна ответить на самый главный вопрос, который делает нас людьми: «Почему?». Я воспринял это как аномалию, поскольку не ожидал, что такие естественные и интуитивные вопросы окажутся вне зоны досягаемости для самых передовых думающих систем.

Только потом я понял, что одна и та же аномалия влияет не только на сферу искусственного интеллекта. Те самые

люди, которых больше всего должно интересовать «Почему?», а именно ученые, трудились в статистической культуре, которая отрицала их право задавать такие вопросы. Конечно, исследователи все равно делали это неформально, но если им хотелось прибегнуть к математическому анализу, приходилось отбрасывать их как ассоциативные.

Изучая эту аномалию, я познакомился с профессионалами из самых разных областей: с философом Кларком Глимором и его коллегами Ричардом Шайнсом и Питером Спиртесом, специалистом по компьютерным наукам Джозефом Халперном, эпидемиологами Джейми Робинсом и Сандером Гренландом, социологом Крисом Уиншипом, статистиками Доном Рубином и Филипом Давидом. Все мы размышляли об одной и той же проблеме и зажгли искру Революции Причинности, которая распространилась, как по цепочке петард, от одной дисциплины к другой и затронула эпидемиологию, психологию, генетику, экологию, геологию, климатологию и т.д. С каждым годом я вижу, что ученые все больше и больше готовы говорить и писать о причинах и следствиях не с извинениями и опущенными глазами, а уверенно и активно. Появилась новая парадигма, в рамках которой основываются утверждения на предположениях, если эти предположения достаточно прозрачны, чтобы вы и другие люди могли судить, насколько они правдоподобны и насколько ваши утверждения чувствительны к их опровержению. Революция Причинности, возможно, не привела к созданию устройства, которое изменило бы нашу жизнь, однако она вызвала трансформацию взглядов, которая неизбежно оздоровит науку.

Я часто думаю, что упомянутая трансформация — второй дар искусственного интеллекта человечеству, и в этой книге в основном рассуждаю об этом. Но сейчас, когда наша история подходит к завершению, пришло время вернуться назад и спросить: в чем же состоит первый дар, для материализации которого потребовалось неожиданно много времени? Приближаемся ли мы к моменту, когда компьютеры или роботы начнут понимать рассуждения о причинно-следственных связях? Способны ли мы создать искусственные интеллекты, не уступающие трехлетним детям в способности воображать?

В этой завершающей главе я не предложу однозначных выводов, но поделюсь соображениями на эту тему.

Каузальные модели и большие данные

За последние годы объемы необработанных данных, которые мы собрали, занимаясь наукой, бизнесом, государственным управлением и даже спортом, вырос в невероятных масштабах. Возможно, эти перемены очевиднее всего тем, кто использует Интернет и социальные сети. Сообщалось, что в 2014 году «Фейсбук» хранил 300 петабайт данных о 2 миллиардах пользователей, или 150 мегабайт данных на каждого пользователя. Игры, в которые играют люди, товары, которые они, вероятно, купят, имена всех их друзей в «Фейсбуке» и, конечно, видео с котиками — все это остается в благословенном океане нулей и единиц.

Распространение огромных баз данных в науке не так очевидно для широкой публики, но не менее важно. Например, для проекта «1 000 геномов» было собрано двести терабайт информации и размещено в так называемом крупнейшем публичном каталоге генетических вариаций. В Архиве космических телескопов имени Барбары Микульски, созданном НАСА, накоплено 2,5 петабайта данных, относящихся к нескольким исследованиям глубокого космоса. Но большие данные повлияли не только на передовую науку, они проникли во все сферы научного знания. Всего одно поколение назад морской биолог мог потратить месяцы, чтобы определить численность любимого вида. Теперь у того же биолога есть моментальный доступ к миллионам единиц информации о рыбе, ее икре, содержимом ее желудка и о чем угодно еще. Вместо того чтобы вести учет, биолог расскажет историю.

Для нас важнее вопрос, что идет дальше. Как извлечь смысл из всех этих чисел, битов и пикселей? Объемы данных могут быть гигантскими, но вопросы мы задаем простые. Этот ли ген вызывает рак легких? В каких солнечных системах вероятнее

встретить планеты, похожие на Землю? Какие факторы сокращают популяцию нашей любимой рыбы и что с этим делать?

В определенных кругах существует почти религиозная вера в то, что ответы на все эти вопросы можно найти в самих данных, если достаточно хорошо провести их интеллектуальный анализ. Однако читатели этой книги обнаружат, что такая страстная убежденность не всегда имеет под собой основания. Вопросы, которые я только что задал, носят каузальный характер, и на них никогда нельзя ответить, ориентируясь только на информацию. От нас требуется подготовить модель процесса, генерирующего данные или, по крайней мере, каких-то ее аспектов. Всякий раз, когда вы видите статью или исследование, где данные анализируют без модели, вы можете быть уверены, что в итоге они будут просто обобщены и, возможно, трансформированы, но не интерпретированы.

Я не хочу сказать, что интеллектуальный анализ данных бесполезен. Он способен стать важнейшим первым шагом, который позволит найти интересные ассоциативные паттерны и точнее поставить вопросы для трактовки. Теперь можно не спрашивать, существуют ли гены, вызывающие рак легких. Вместо этого достаточно просканировать геном и найти гены, у которых высокая корреляция с раком легких (как в примере с Большим, описанном в главе 9). Потом ставится вопрос, вызывает ли этот ген рак легких (и каким образом). Мы никогда бы не задали вопрос об этом гене, если бы у нас не было интеллектуального анализа данных. Однако, чтобы пойти дальше, необходимо разработать причинную модель, уточняющую, например, на какие переменные воздействует ген, какие здесь возможны осложнители и какие иные причинные пути способны достичь этого результата. Интерпретация данных подразумевает выдвижение гипотез о том, как все происходит в реальном мире.

Еще одна роль больших данных в задачах для причинного вывода открывается в механизме причинного анализа, описанном во вступлении, на последнем этапе его работы (шаг 8), где мы переходим от оцениваемой величины к оценке. Эта стадия статистической оценки принимает нетривиальный

оборот, когда число переменных велико, и только современные технологии интеллектуального анализа данных и машинного обучения помогут справиться с проклятием многомерности. Это объясняет, почему большие данные и причинный вывод сегодня играют важнейшую роль в развивающейся персонализованной медицине. Здесь мы пытаемся извлечь выводы из прошлого поведения группы индивидов, как можно более похожих по максимальному набору параметров на данного конкретного человека. С помощью причинного вывода мы отбрасываем нерелевантные характеристики и извлекаем этих индивидов из разнообразных исследований, в то время как большие данные позволяют собрать о них достаточно информации.

Легко понять, почему некоторые люди считают интеллектуальный анализ данных финальным, а не первым шагом. Он обещает решение с использованием имеющихся технологий. Он избавляет и нас, и машины будущего от необходимости рассматривать и формулировать обоснованные предположения о том, как устроен мир. В некоторых областях наши знания находятся в таком зачаточном состоянии, что мы понятия не имеем, как приступить к созданию модели мира. Но большие данные не решат эту проблему. Важнейшая часть ответа должна исходить из модели, нарисованной нами или предложенной и уточненной машинами.

Чтобы не показаться излишне критичным по отношению к работе с большими данными, я хотел бы упомянуть одну новую возможность для их симбиоза с причинным выводом. Она называется транспортабельностью.

Благодаря большим данным мы можем получить доступ к огромному количеству не только людей в любом конкретном эксперименте, но и исследований, проведенных в разных местах и в различных условиях. Часто нам нужно объединить результаты этих исследований и перенести их на новые группы населения, которые могут отличаться даже в том, что будет для нас неожиданным.

Процесс перевода результатов исследования из одних условий в другие играет в науке фундаментальную роль. Фактиче-

ски научный прогресс остановился бы, если бы у нас не было способности обобщать результаты лабораторных экспериментов и переносить их в реальный мир, например из пробирок на животных и на людей. Но до недавнего времени каждой науке приходилось разрабатывать собственные критерии для отделения валидных обобщений от невалидных, а систематических методов для решения проблемы транспортабельности в целом не существовало.

За последние пять лет мне и моему бывшему студенту (теперь коллеге) Элиасу Барейнбойму удалось найти исчерпывающий критерий, чтобы принять решение о том, переносимы ли результаты. Как обычно, необходимое условие для его использования — представить процесс генерации данных в виде диаграммы причинности, на которой отмечены места потенциальных несоответствий. Переносить результат не обязательно означает принимать его в исходной форме и применять в новой среде. Исследователю, возможно, придется откалибровать его, чтобы учесть различия между двумя средами.

Предположим, мы хотим узнать эффект воздействия рекламы в Интернете (X) на вероятность того, что потребитель купит товар (Y), скажем доску для серфинга. У нас есть данные, полученные в результате исследований в пяти разных местах — в Лос-Анджелесе, Бостоне, Сан-Франциско, Торонто и Гонолулу. Теперь мы хотим оценить, насколько эффективной эта реклама будет в Арканзасе. К сожалению, все группы и все исследования несколько отличаются. Например, группа, изученная в Лос-Анджелесе, моложе, чем наша целевая аудитория, а в Сан-Франциско она отличается по количеству переходов по ссылке. На рис. 65 показаны уникальные характеристики каждой группы и каждого исследования. Можем ли мы объединить данные, полученные в далеких друг от друга местах, чтобы оценить эффективность рекламы в Арканзасе? Можем ли мы сделать это, не собрав данные в Арканзасе? Или измерив лишь ограниченное число переменных? Или проведя пилотное наблюдательное исследование?

Рис. 65. Проблема транспортабельности

На рис. 66 эти различия переведены в форму графика. Переменная *Z* представляет возраст, который играет роль осложнителя; молодые люди с большей вероятностью увидят рекламу и с большей вероятностью купят продукт, даже если не видели рекламу. Переменная *W* отражает переход по ссылке с целью получить дополнительную информацию. Это медиатор — шаг, который необходим, чтобы просмотр рекламы превратился в покупку продукта. Буква *S* в каждом случае обозначает переменную, «производящую различие», т.е. гипотетическую переменную, которая указывает на характеристики, отличающие две группы. Например, в группе *б* «Лос-Анджелес» индикатор *S* указывает на *Z*, возраст. В каждом из иных городов индикатор указывает на характерную черту группы, приведенную на рис. 65

Для рекламного агентства хорошая новость здесь в том, что компьютер теперь способен справиться с этой сложной проблемой слияния данных и, руководствуясь *do*-исчислением, сообщить нам, какие исследования используются для ответа на наш запрос и какими способом это делается, а также какую информацию нам нужно собрать в Арканзасе, чтобы подтвердить вывод. В некоторых случаях эффект переносится напрямую, без дополнительной работы — возможно, нам не придется ехать в Арканзас. Например, эффект от рекламы в Арканзасе должен быть таким же, как в Бостоне, потому что

согласно диаграмме, группа с отличается от группы а только переменной V, которая не влияет ни на воздействие *X*, ни на результат *Y*.

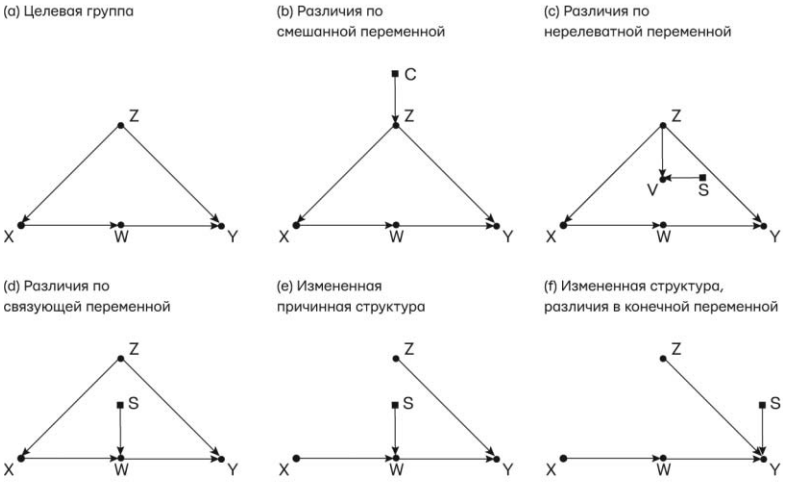


Рис. 66. Различия между исследованными группами, выраженные в графической форме

Нам необходимо по-новому оценить данные в некоторых других исследованиях, положим, принять в расчет иную возрастную структуру населения в лос-анджелесском исследовании *б*. Интересно, что эксперимента в Торонто *е* достаточно для оценки нашего запроса в Арканзасе, несмотря на несоответствие в параметре *W*, если мы можем измерить только *X*, *W* и *Y* в Арканзасе.

Примечательно, что мы нашли примеры, в которых транспортировка невозможна из любого отдельно взятого исследования; тем не менее целевое количество можно оценить по их комбинации. Кроме того, даже исследования, откуда нельзя ничего перенести, не совсем бесполезны. Так, исследование

Гонолулу *e* на рис. 66 невозможно транспортировать из-за стрелки S → *Y*. Однако стрелка *X* → *W* не загрязнена *S*, поэтому данные, полученные в этой группе, можно использовать для оценки *P* (*W* | *X*). Объединив это с оценками *P* (*W* | *X*) из других исследований, мы повысим точность этого подвыражения. Тщательно комбинируя такие подвыражения, мы можем синтезировать точную общую оценку целевого количества.

Хотя в простых случаях эти результаты интуитивно разумны, когда диаграммы становятся более сложными, нам нужна помощь формального метода. *Do*-исчисление обеспечивает общий критерий для определения транспортабельности в таких случаях. Правило довольно простое: если выполняется допустимая последовательность *do*-операций (с использованием правила из главы 7), которые преобразуют целевую величину в другое выражение, в котором любой фактор, включающий *S*, не содержит *do*-операторов, тогда оценка транспортабельна. Логика проста: любой такой фактор оценивается по имеющимся данным, не затронутым фактором несоответствия *S*.

Элиас Баренбойм сумел сделать с проблемой транспортабельности то же, что Илья Шпицер совершил с проблемой интервенции. Он разработал алгоритм, который автоматически определяет, является ли желаемый эффект переносимым, используя только графические критерии. Другими словами, он сообщает, реально ли отделить S от *do*-операторов или нет.

Результаты Барейнбойма впечатляют, потому что в их свете явление, которое раньше считалось угрозой для валидности, превратилось в новую возможность. Она позволяет нам применять многочисленные исследования, для участников которых нельзя установить собственные критерии (и, соответственно, гарантировать, что исследуемая группа окажется такой же, как интересующая) в наших целях. Вместо того чтобы рассматривать эту разницу как угрозу для внешней валидности исследования, теперь мы устанавливаем валидность в ситуациях, которые раньше казались безнадежными. Именно потому, что мы живем в эпоху больших данных, у нас есть доступ к информации о многих исследованиях и вспомогательных

переменных (например, *Z* и *W*), которые позволят нам переносить результаты с одной группы на другую.

Попутно упомяну, что Барейнбойм также подтвердил аналогичные результаты для другой проблемы, которая долгое время беспокоила статистиков, — систематической ошибки отбора. Этот вид ошибки возникает, когда изучаемая группа отличается от целевой по какому-либо значимому признаку, что весьма похоже на проблему транспортабельности. Да, эти явления действительно похожи, за исключением одного очень важного отличия: вместо того чтобы рисовать стрелку от индикаторной переменной *S* к затронутой переменной, мы рисуем стрелку в направлении *S*. Пусть *S* обозначает отбор (участников исследования). Скажем, если в нашем исследовании участвовали бы только госпитализированные пациенты, как в примере с ошибкой Берксона, мы бы нарисовали стрелку от госпитализации к *S*, показывая, что именно она является причиной отбора для нашего исследования. В главе 6 мы рассматривали эту ситуацию только как угрозу для валидности эксперимента. Но теперь получится воспринимать ее как возможность. Если мы поймем механизм, с помощью которого отбираются испытуемые, то преодолеем систематическую ошибку, собрав данные о правильном наборе упростителей и применив соответствующую формулу повторного взвешивания или корректировки. Работа Барейнбойма позволяет нам использовать каузальную логику и большие данные, чтобы творить чудеса, которые раньше были немыслимы.

Слова «чудеса» и «немыслимы» редко встречаются в научном дискурсе, и читатель может задаться вопросом, не слишком ли много у меня энтузиазма. Но у меня есть достойная причина, чтобы высказаться именно так. Концепция внешней достоверности как угрозы экспериментальной науке существует уже по крайней мере полвека, с тех пор как Дональд Кэмпбелл и Джулиан Стэнли признали и дали определение этого термина в 1963 году. Я разговаривал с десятками специалистов и выдающихся авторов, которые писали на эту тему. К моему удивлению, ни один из них не смог решить ни одну из «игрушечных задач», представленных на рис. 66. Я называю

их игрушечными, потому что их легко описать, легко решить и легко проверить.

В настоящее время культура внешней валидности полностью сосредоточена на перечислении и категоризации угроз, а не на борьбе с ними. Более того, она настолько парализована угрозами, что сама идея нейтрализации угроз встречается с подозрением и недоверием. Специалистам, которые плохо разбираются в графических моделях, проще учесть дополнительные угрозы, чем попытаться устранить любую из них. Я надеюсь, что такие слова, как «чудеса», должны сподвигнуть моих коллег рассматривать подобные проблемы как интеллектуальные вызовы, а не причины погрузиться в отчаяние.

Я хотел бы представить читателю отчеты о случаях, когда удалось успешно справиться с задачами транспортировки и преодолеть систематическую ошибку отбора, но эти методы все еще слишком новы, чтобы получить широкое применение. Тем не менее я твердо уверен в том, что исследователи вскоре откроют для себя силу алгоритмов Барейнбойма, и тогда внешняя валидность, как и проблема осложнителей до того, утратит мистическую и устрашающую силу.

Сильный ИИ и свобода воли

Еще не просохли чернила в великом тексте Тьюринга «Вычислительные машины и разум», как научные фантасты и футурологи уже начали оценивать перспективы, связанные с думающими машинами. Порой они представляли эти машины как безвредных или даже благородных персонажей — вроде жужжащего и щебечущего R2-D2 и андроида с британскими манерами C-3PO из «Звездных войн». Но порой машины оказывались гораздо более зловещими и даже готовились уничтожить человечество, как в «Терминаторе», или поработить людей, заключив их в виртуальную реальность, как в «Матрице».

Во всех этих случаях представление об ИИ больше говорит о тревогах сценаристов или о возможностях отдела спецэффектов, чем о реальных исследованиях искусственного интеллек-

та. Создать его оказалось гораздо труднее, чем предполагал Тьюринг, даже несмотря на то, что чистая вычислительная мощность наших компьютеров, несомненно, превзошла его ожидания.

В главе 3 я описал некоторые причины такого медленного прогресса. В 1970-е и начале 1980-х исследованиям в области искусственного интеллекта мешала концентрация на системах, основанных на правилах. Но такие системы оказались неверным путем. Они были очень хрупкими. Любое небольшое изменение в их рабочих допущениях требовало переписывания программы. Они не могли справиться с неопределенностью или противоречивыми данными. Наконец, они не были прозрачными с научной точки зрения; нельзя было математически доказать, что они будут вести себя определенным образом, и нельзя было определить, что ремонтировать, когда они этого не делали. Не все исследователи ИИ возражали против отсутствия прозрачности. В то время в этом направлении появилось разделение на «аккуратистов» (тех, кто хотел видеть прозрачные системы с гарантированным поведением) и «нерях» (тех, кто просто хотел, чтобы системы работали). Я всегда был «аккуратистом».

Мне посчастливилось прийти в эту сферу в тот момент, когда все было готово к новому подходу. Байесовские сети были вероятностными; они могли справиться с миром, полным противоречивых и неопределенных данных. В отличие от систем, основанных на правилах, они были модульными и легко внедрялись на платформе распределенных вычислений, что обеспечивало быструю работу. Наконец, для меня (и других «аккуратистов») было важно, что байесовские сети работали с вероятностями математически надежным способом, т.е. мы знали: если что-то шло не так, ошибка была в программе, а не в наших рассуждениях.

Но даже со всеми этими преимуществами байесовские сети все еще не понимали причинно-следственных связей. Они устроены так, что информация там течет в обоих направлениях, причинном и диагностическом: дым увеличивает вероятность возгорания, а пожар — вероятность возникновения

дыма. Фактически байесовская сеть не способна отличить «причинное направление». Погнавшись за этой аномалией — чудесной аномалией, как выяснилось потом, — я отвлекся от машинного обучения и перешел к изучению причинности. Я не мог смириться с мыслью, что будущие роботы не смогут общаться с нами на нашем родном причинно-следственном языке. Оказавшись в стране причинности, я, естественно, увлекся обширным спектром других наук, в которых каузальная асимметрия имеет первостепенное значение.

И вот в последние 25 лет я держался вдали от родной страны автоматизированного мышления и машинного обучения. Тем не менее издалека мне хорошо видны современные тенденции и модные направления.

В последние годы наиболее заметный прогресс в области ИИ связан с так называемым глубоким обучением, в котором используются такие методы, как сверточные нейронные сети. Эти сети не следуют правилам вероятности; они не решают проблему неопределенности ни строго, ни прозрачно. И еще в меньшей степени они подразумевают сколько-нибудь явное представление среды, в которой действуют. Вместо этого архитектура сети способна развиваться сама по себе. Закончив обучение новой сети, программист понятия не имеет, какие вычисления она выполняет и почему они работают. Если сеть выходит из строя, непонятно, как это исправить.

Возможно, прототипическим примером здесь будет AlphaGo, программа на основе сверточной нейронной сети, которая играет в древнюю азиатскую игру го. Ее разработала DeepMind, дочерняя компания Google. Го всегда считалась самой трудной для ИИ среди всех человеческих игр с полной информацией. Хотя компьютеры обыграли людей в шахматы еще в 1997 году, даже в 2015 году они еще не могли тягаться с профессиональными игроками самого низкого уровня. Сообщество игроков в го считало, что до настоящего сражения людей с компьютерами должны пройти десятилетия.

Это изменилось почти в мгновение ока с появлением AlphaGo. Большинство игроков в го впервые услышали о программе в конце 2015 года, когда она победила человека-про-

фессионала со счетом 5:0. В марте 2016 года AlphaGo выиграла у Ли Седола, долгие годы считавшегося сильнейшим игроком среди людей, со счетом 4:1. Через несколько месяцев программа провела 60 онлайн-игр с лучшими игроками, не проиграв ни одной, а в 2017 году официально завершила карьеру после победы над действующим чемпионом мира Ке Цзе. Партия, проигранная Седолу, так и останется единственной, в которой она уступила человеку.

Все это очень впечатляет, и результаты не оставляют сомнений: глубокое обучение работает для определенных задач. Но это полная противоположность прозрачности. Даже программисты AlphaGo не могут сказать вам, почему эта программа играет так хорошо. По опыту они знали, что глубокие сети успешно решают задачи компьютерного зрения и распознавания речи. В то же время наше понимание глубокого обучения полностью эмпирическое и не дает никаких гарантий. В начале команда AlphaGo не могла предвидеть, что программа победит лучшего из игроков-людей через год, два или пять лет. Они просто экспериментировали, и все получилось.

Кто-то скажет, что прозрачность на самом деле не нужна. У нас нет детального понимания того, как работает человеческий мозг, и все же он работает хорошо, и мы прощаем себе эту скудость понимания. В таком случае, почему не использовать системы глубокого обучения, чтобы создать новый вид интеллекта, не понимая, как он работает? Я не могу утверждать, что это неверный подход. В настоящий момент «неряхи» взяли на себя инициативу. Тем не менее скажу, что лично мне не нравятся непрозрачные системы, и поэтому я не собираюсь их исследовать.

Оставив в стороне мои предпочтения, к аналогии с человеческим мозгом можно добавить еще один фактор. Да, мы прощаем себе скудное понимание работы человеческого мозга, но все еще можем общаться с другими людьми, учиться у них, наставлять их и мотивировать на нашем родном языке причин и следствий. Мы делаем это, потому что наш мозг работает так же, как и у них. Но если все наши роботы будут такими же

непрозрачными, как AlphaGo, мы не сможем вести с ними содержательные разговоры, а это будет весьма прискорбно.

Когда мой домашний робот включает пылесос, пока я еще сплю, и я говорю ему: «Не надо было меня будить», я хочу, чтобы он понял: не стоило пылесосить. Но я не хочу, чтобы он интерпретировал эту жалобу как указание никогда больше не пылесосить наверху. Он должен понимать то, что прекрасно понимаем мы с вами: пылесосы шумят, шум будит людей и некоторые люди этому не рады. Другими словами, наш робот должен понимать причинно-следственные связи — по сути, контрфактивные отношения вроде закодированных во фразе «не надо было».

Действительно, обратите внимание на богатое содержание этого короткого предложения с инструкцией. Нам не нужно сообщать роботу, что то же самое относится к уборке пылесосом внизу или где-либо еще в доме, но не когда я бодрствую или отсутствую и не в случае, если пылесос оснащен глушителем и т.д. Может ли программа глубокого обучения понять всю полноту этой инструкции? Вот почему я не удовлетворен очевидно прекрасной производительностью непрозрачных систем. Прозрачность обеспечивает эффективное общение.

Но один аспект глубокого обучения меня все-таки интересует: теоретические ограничения этих систем и, в первую очередь, ограничения, проистекающие из их неспособности выйти за пределы первого уровня на Лестнице Причинности. Это ограничение не препятствует работе AlphaGo в узком мире игры го, поскольку описание доски вместе с правилами игры составляет адекватную причинную модель для мира го. Тем не менее это препятствует системам обучения, которые действуют в средах, управляемых насыщенными сетями причинных сил, но имея при этом доступ только к поверхностным их проявлениям. Медицина, экономика, образование, климатология и социальная сфера — типичные примеры таких сред. Подобно узникам в знаменитой пещере Платона, системы глубокого обучения исследуют тени на стене и учатся точно предсказывать их движения. Им не хватает понимания того, что наблюдаемые тени — лишь проекции трехмерных

объектов, движущихся в трехмерном пространстве. Сильный ИИ требует этого понимания.

Исследователи глубокого обучения знают об этих основных ограничениях. Так, экономисты, использующие машинное обучение, отметили, что их методы не отвечают на ключевые вопросы нынешнего времени, положим не позволяют оценить, как подействуют неопробованные пока методы и меры. Типичные примеры здесь — новые принципы ценообразования, субсидии, изменение минимальной заработной платы. С технической точки зрения методы машинного обучения сегодня обеспечивают эффективный способ перейти от анализа конечных выборок к распределениям вероятностей, но нам еще только предстоит перейти от последних к причинно-следственным связям.

Когда мы начинаем говорить о сильном ИИ, причинные модели превращаются из роскоши в необходимость. Для меня сильный ИИ — это машина, которая может размышлять о своих действиях и извлекать уроки из совершенных ошибок. Она должна понимать высказывание «Надо было поступить иначе» независимо от того, говорит ли это ей человек или она сама приходит к такому выводу. Контрфактивная интерпретация этого утверждения выглядит так: «Я сделал $X = X$, и результат был $Y = Y$. Но если бы я действовал иначе, скажем, $X = X¢$, то результат был бы лучше, возможно, $Y = Y¢$». Как мы уже увидели, оценка таких вероятностей была полностью автоматизирована при наличии достаточного объема данных и адекватно обозначенной причинной модели.

Более того, я думаю, что очень важной целью для машинного обучения будет более простая вероятность $P\ (Y_X = X_1 = Y¢ \mid X = X)$, когда машина наблюдает $X = X$, но не результат Y, а затем спрашивает о результате альтернативного события $X = X¢$. Если машина способна вычислить эту величину, то это преднамеренное действие можно рассмотреть как наблюдаемое событие $(X = x)$ и спросить: «А если я поменяю решение и сделаю вместо этого $X = X¢$?» Это выражение математически эквивалентно эффекту лечения на уже пролеченных (упо-

минается в главе 8), и у нас есть масса результатов, которые показывают, как его оценить.

Намерение — очень важная составляющая в процессе принятия решений. Если бывший курильщик чувствует позыв зажечь сигарету, ему стоит очень хорошо подумать о том, какие причины стоят за этим намерением и спросить, не приведет ли противоположное действие к лучшему исходу. Способность формировать собственное намерение и использовать его как доказательство в причинно-следственных рассуждениях — это уровень осознанности (если не самосознания), которого не достигла ни одна из известных мне машин. Я хотел бы иметь возможность ввести машину в искушение и увидеть, как она говорит: «Нет».

Обсуждение намерения неизбежно ведет к еще одной проблеме, стоящей перед сильным искусственным интеллектом, — проблеме свободы воли. Если мы попросим машину заиметь намерение сделать $X = x$, осознать его и выбрать вместо этого $X = x'$, то представляется, что мы попросим ее проявить свободную волю. Но как у робота может быть свобода воли, если он просто следует инструкциям, записанным в программе?

Философ Джон Сёрл из Калифорнийского университета в Беркли назвал проблему свободы воли «скандалом в философии» отчасти потому, что со времен античности не произошло никакого прогресса в ее разработке, и отчасти потому, что мы не можем отмахнуться от нее как от оптической иллюзии. Вся наша концепция «себя» предполагает присутствие такой вещи, как свобода выбора. Например, нет никакого очевидного способа примирить мое яркое, безошибочное ощущение возможности (скажем, прикоснуться или не прикоснуться к носу) с пониманием реальности, которая предполагает каузальный детерминизм: все наши действия вызваны электрическими нейронными сигналами, идущими от мозга.

В то время как многие философские проблемы со временем исчезли в свете научного прогресса, свободная воля упрямо остается загадкой, такой же свежей, какой она казалось Аристотелю и Маймониду. Более того, хотя основания для свободного волеизъявления порой находили в сфере духовности или

теологии, эти объяснения не подошли бы для программируемой машины. Поэтому появление свободной воли у роботов непременно будет результатом ухищрений — по крайней мере, такова догма.

Не все философы считают, что между свободной волей и детерминизмом действительно есть столкновение. Группа под названием «совпаденцы», к которой я причисляю с себя, считает это лишь очевидным столкновением между двумя уровнями описания: нейронным, на котором процессы кажутся детерминистскими (за исключением квантового индетерминизма), и когнитивным, на котором у нас есть живое ощущение возможностей. Такие явные несоответствия не столь уж редки в науке. Например, уравнения физики обратимы во времени на микроскопическом уровне, но кажутся необратимыми на макроскопическом уровне описания; дым никогда не возвращается в дымоход. Но это поднимает новые вопросы: если допустить, что свобода воли является (или может быть) иллюзией, почему для нас, людей, так важно иметь эту иллюзию? Почему эволюция приложила усилия, чтобы наделить нас этой концепцией? Уловка это или нет, должны ли мы запрограммировать следующее поколение компьютеров, чтобы создать эту иллюзию? Зачем? Какие вычислительные преимущества это влечет за собой?

Я думаю, что для решения этой вечной загадки — проблемы примирения свободы воли с детерминизмом — нужно понять преимущества иллюзии свободы. Проблема исчезнет на наших глазах, если мы наделим детерминированную машину такими же преимуществами.

Вместе с этой загадкой о функции мы также должны справиться с задачей симуляции. Если нейронные сигналы от мозга запускают все наши действия, то перед ним стоит вечная задача постоянно определять одни действия как «волевые» или «намеренные», а другие — как «непреднамеренные». Что именно представляет собой этот процесс маркировки? Какая нейронная связь обеспечит тому или иному сигналу звание «намеренного»? Во многих случаях добровольные действия распознаются по следу, который они оставляют в кратковремен-

ной памяти, причем этот след отражает цель или мотивацию. Например: «Почему ты это сделала?» — «Хотела произвести на тебя впечатление». Или, как невинно ответила Ева: «Змей обольстил меня, и я ела». Но во многих других случаях намеренное действие совершается, однако у него нет очевидных причин или мотивов. Рационализация действий может быть реконструктивным процессом, который происходит уже после действия. Так, футболист может объяснить, почему он решил передать мяч Джо, а не Чарли, но редко бывает, чтобы эти причины сознательно запускали действие. В пылу игры тысячи входных сигналов соревнуются за внимание игрока. Важнейшее решение состоит в том, какому сигналу дать приоритет, а причины трудно вспомнить и сформулировать.

Таким образом, все исследователи искусственного интеллекта пытаются ответить на два вопроса — о функции и симуляции, причем первый ведет за собой второй. Как только мы поймем, какую вычислительную функцию выполняет свободная воля в нашей жизни, мы приступим к оснащению машин такими же функциями. Это станет инженерной проблемой, хотя и весьма сложной.

Для меня четко выделяются некоторые аспекты функционального вопроса. Иллюзия свободы воли дает нам возможность говорить о намерениях и обдумывать их рационально, вероятно, используя контрфактивную логику. Скажем, тренер удаляет нас с футбольного матча и говорит: «Надо было пасовать Чарли». Подумайте обо всех сложных значениях, заключенных в этих восьми словах.

Во-первых, цель этого указания «надо было» — быстро передать ценную информацию от тренера игроку: в будущем, столкнувшись с подобной ситуацией, выбирайте действие *B*, а не действие *A*. Но похожих ситуаций слишком много, чтобы их перечислить, и они вряд ли известны даже самому тренеру. Вместо того чтобы перечислять особенности этих похожих ситуаций, тренер указывает на действия игрока, которые отражают его намерения во время принятия решения. Объявляя действие неудачным, тренер просит игрока определить программу, которая привела к его решению, а затем переустановить в ней

приоритет, чтобы «пасовать Чарли» стало предпочтительным действием. В этой инструкции есть глубокая мудрость, потому что кто, если не сам игрок, знает свою программу? Это безымянные нейронные пути, которые неведомы тренеру или любому внешнему наблюдателю. Просьба к игроку совершить действие, отличное от предпринятого, означает, что поощряется анализ намерений, подобный описанному выше. Таким образом, обдумывание намерений позволяет преобразовать сложные каузальные указания в простые.

Соответственно, я бы предположил, что команда роботов лучше бы играла в футбол, если бы их запрограммировали общаться так, словно у них есть свобода воли. Независимо от того, насколько техничны отдельные роботы, результативность их команды улучшится, если они смогут разговаривать друг с другом, как если бы они были не заранее запрограммированными роботами, а автономными агентами, полагающими, что у них есть выбор.

Хотя пока неизвестно, улучшит ли иллюзия свободы воли взаимодействие роботов с роботами, в плане взаимодействия роботов с людьми неуверенности гораздо меньше. Для естественного общения с людьми сильному ИИ, безусловно, потребуется понимать вокабуляр возможностей и намерений, и, следовательно, имитировать иллюзию свободной воли. Как я объяснил выше, им также выгодно самим «верить» в свою свободную волю — чтобы иметь возможность осознать собственное намерение и изменить действие.

Способность рассуждать о своих убеждениях, намерениях и желаниях представляет серьезный вызов для исследователей ИИ и определяет понятие свободы действий. Однако философы изучают эту способность как часть классической проблемы самосознания. Такие вопросы, как «Могут ли машины иметь сознание?» или «Чем программа со свободой действий отличается от обычной программы?», занимали и занимают лучшие умы многих поколений, и я не буду притворяться, что готов дать на них полные ответы. Тем не менее я считаю, что алгоритмизация контрфактивного — важный шаг к пониманию этих вопросов и к превращению сознания и действия в вы-

числительную реальность. Методы, описанные для оснащения машины символическим представлением окружающей среды и способностью представить гипотетическое возмущение этой среды, могут быть расширены, чтобы включить в среду саму машину. Ни одна машина не способна обработать полную копию собственного программного обеспечения, однако ей можно дать схему его основных компонентов. Другие компоненты затем станут рассуждать об этой схеме и имитировать состояние самосознания.

Чтобы создать ощущение свободы действий, мы также должны оснастить этот программный пакет памятью для записи прошлых активаций, на которые он будет ссылаться, когда его спросят: «Почему ты это сделал?» Действия, которые проходят через определенные паттерны при активации пути, получат аргументированные объяснения, например: «Потому что альтернатива оказалась менее привлекательной». Другие ответы будут бесполезными или уклончивыми: «Хотел бы я знать почему» или «Потому что ты меня так запрограммировал».

Подводя итог, я считаю, что программный пакет, который даст мыслящей машине преимущества свободы воли, будет состоять как минимум из трех частей: из причинной модели мира, причинной модели собственного программного обеспечения, пусть и поверхностной, и памяти, в которой будет записано, как намерения в ее уме соответствуют событиям во внешнем мире.

Возможно, именно мы сами начинаем понимать причины в раннем детстве. У нас в сознании присутствует что-то вроде генератора намерений, который говорит нам, что мы должны предпринять действие $X = x$. Но дети любят экспериментировать — бросать вызов родителям, учителям и даже собственным изначальным намерениям — и делать что-то новое, просто для удовольствия. Полностью осознавая, что мы должны сделать $X = x$, мы игриво делаем $X = x¢$. Потом мы смотрим, что происходит, повторяем процесс и ведем отчетность о том, насколько хорош наш генератор намерений. Наконец, когда мы начинаем настраивать собственное программное обеспечение, именно тогда мы и берем на себя моральную ответственность за свои

действия. Эта ответственность выступает иллюзией на уровне нейронной активации, но не на уровне программного обеспечения самосознания.

Вдохновленный этими возможностями, я считаю, что сильный ИИ с пониманием причин и способностями к самостоятельным действиям — это реально. И отсюда вытекает вопрос, который писатели-фантасты задают с 1950-х годов: стоит ли нам волноваться? Является ли сильный ИИ ящиком Пандоры, который мы не должны открывать?

Недавно публичные персоны вроде Илона Маска и Стивена Хокинга заявили, что у нас есть причины беспокоиться. Маск написал в «Твиттере», что ИИ «потенциально опаснее ядерного оружия». В 2015 году на веб-сайте Джона Брокмана Edge.org был задан ежегодный вопрос, и он выглядел так: «Что вы думаете о машинах, которые думают?». На него дали 186 вдумчивых и провокационных ответов, которые потом были собраны в книге под названием «Что думать о машинах, которые думают» *(What to Think About Machines That Think)*.

Намеренно расплывчатый вопрос Брокмана подразделяется как минимум на пять связанных между собой:

1. Мы уже создали мыслящие машины?
2. Можем ли мы создать мыслящие машины?
3. Будем ли мы создавать мыслящие машины?
4. Стоит ли создавать мыслящие машины?

И наконец, незаданный вопрос, который лежит в основе наших тревог:

5. Можем ли мы создать машины, способные отличать добро и зло?

Ответ на первый вопрос будет отрицательным, но я полагаю, что на все остальные можно ответить утвердительно. Мы точно еще не создали машины, которые думают, интерпретируя мир как люди. До этого момента реально всего лишь симулировать человеческое мышление в узко определенных областях, где есть только самые примитивные каузальные структуры. Для них действительно удается создавать машины, которые превос-

ходят людей, но это не должно удивлять, ведь в этих областях вознаграждается единственная вещь, которую компьютеры делают хорошо, а именно вычисления.

Ответ на второй вопрос почти точно положительный, если определять способность думать как способность пройти тест Тьюринга. Я утверждаю это на основании того, что мы извлекли из мини-теста Тьюринга. Умение отвечать на запросы на всех трех уровнях Лестницы Причинности создает основу для «самосознающего» ПО, которое позволяет машине думать о своих намерениях и размышлять о своих ошибках. Алгоритмы для ответа на каузальные и контрфактивные запросы уже существуют (во многом благодаря моим студентом), и чтобы их внедрить, не хватает только трудолюбивых исследователей искусственного интеллекта.

Третий вопрос, конечно, зависит от событий в человеческой истории, которые трудно предсказать. Но исторически люди редко отказывались от того, чтобы создавать или практиковать вещи, если это позволяет развитие техники. Отчасти это объясняется тем, что мы не в состоянии узнать, на что способны наши технологии, пока не используем их на практике, будь то клонирование животных или отправка людей на Луну. Однако поворотным пунктом в этом процессе стал взрыв атомной бомбы: многие люди считают, что эту технологию лучше было бы не создавать.

Хороший пример того, как ученые отказались от вполне возможного на практике в послевоенный период, подала Асиломарская конференция 1975 года о рекомбинации ДНК — новой технологии, которая освещалась в СМИ в весьма апокалиптических тонах. Ученым, работающим в этой области, удалось прийти к консенсусу относительно разумных методов обеспечения безопасности, и достигнутое ими тогда согласие сохранялось в течение следующих четырех десятилетий. Сейчас рекомбинантная ДНК — обычная зрелая технология.

В 2017 году Институт будущего жизни (*Future of Life Institute*) организовал в Асиломаре похожую конференцию об искусственном интеллекте, где на месте было достигнуто соглашение о 23 принципах для будущих исследований «полезного ИИ».

Хотя большинство из собранных в нем указаний не относится к темам, обсуждаемым в этой книге, рекомендации по этике и ценностям определенно достойны внимания. Например, рекомендация 6: «Системы ИИ должны быть безопасными и надежными в течение всего срока их эксплуатации, и это должно быть верифицируемо». Или рекомендация 7: «Если система ИИ причиняет вред, должна существовать возможность установить причину», — из которой ясно, как важна прозрачность. Рекомендация 10: «Системы ИИ с высокой степенью автономности должны быть разработаны таким образом, чтобы их цели и поведение были согласованы с человеческими ценностями на всем протяжении работы» — в предложенной формулировке выглядит довольно расплывчато, но ей можно было бы придать смысл на практике, если бы от этих систем требовалась способность заявлять о намерениях и сообщать людям о причинах и следствиях.

Мой ответ на четвертый вопрос тоже положительный, он основан на ответе на пятый вопрос. Я верю, что мы научимся создавать машины, которые смогут отличать добро от зла, по крайней мере так же надежно как люди и, надеюсь, даже лучше. Первое требование к этичной машине — способность размышлять над собственными действиями, что подпадает под анализ контрфактивных суждений. Как только мы запрограммируем самосознание, пусть и ограниченное, эмпатия и справедливость последуют за ним, поскольку они основаны на тех же вычислительных принципах, когда в уравнение добавляется еще один агент.

Между причинным подходом к созданию этичного робота и подходом, который изучается и постоянно пересказывается в научной фантастике с 1950-х годов — законами робототехники Азимова, — существует принципиальное различие. Айзек Азимов предложил три абсолютных закона, начиная со следующего: «Робот не может причинить вред человеку или своим бездействием допустить, чтобы человеку был причинен вред». Но, как неоднократно доказывала научная фантастика, законы Азимова всегда приводят к противоречиям. Для ученых, занимающихся искусственным интеллектом, это неудивительно:

системы, основанные на правилах, никогда не работают без сбоев. Но отсюда не следует, что создать морального робота невозможно. Просто подход здесь не может быть предписывающим и основанным на правилах. Отсюда следует, что мы должны обеспечить мыслящие машины теми же когнитивными способностями, которые есть у нас, включая эмпатию, долгосрочное прогнозирование и сдержанность, а затем позволить им принимать собственные решения.

Как только мы создадим робота с моральными принципами, апокалиптические видения постепенно перестанут нас пугать. Нет причины отказываться от создания машин, которые будут лучше нас отличать добро и зло, сопротивляться искушениям, признавать вину и заслуги. На этом этапе мы, подобно шахматистам и игрокам в го, даже начнем учиться у собственных творений. Мы сможем положиться на наши машины, когда нам понадобится ясно и причинно обусловленное чувство справедливости. Мы поймем, как работает наше собственное программное обеспечение, обеспечивающее свободу воли, и как ему удается скрывать от нас свои секреты. Такая мыслящая машина была бы прекрасным спутником для нашего биологического вида и действительно могла бы считаться первым и лучшим подарком человечеству со стороны искусственного интеллекта.

Благодарности

Перечислять всех учеников, друзей, коллег и учителей, которые внесли свой вклад в эту книгу, было бы равносильно написанию еще одной книги. Тем не менее, несколько игроков заслуживают особого упоминания, с моей личной точки зрения. Я хотел бы поблагодарить Фила Давида за то, что он дал мне мое первое прослушивание на страницах *B*iometrika; Джейми Робинса и Сандера Гренланди за то, что превратили эпидемиологию в сообщество, говорящее на графиках; покойного Денниса Линдли за то, что он дал мне уверенность в том, что даже опытные статистики могут распознать недостатки в своей области и сплотиться для ее улучшения; Крису Уиншипу, Стивену Моргану и Феликсу Элверту за вступление социальной науки в эпоху причинности; и, наконец, Питеру Спиртсу, Кларку Глимуру и Ричарду Шайнсу за их помощь в том, чтобы столкнуть меня с утеса вероятностей в бурные воды причинности.

Углубляясь в свою древнюю историю, я должен поблагодарить Джозефа Хермони, доктора Шимшона Ланге, профессора Франца Оллендорфа и других преданных своему делу учителей естественных наук, которые вдохновляли меня от начальной школы до колледжа. Они привили многим из нас, израильтян первого поколения, чувство миссии и исторической ответ-

ственности за проведение научных исследований как самой благородной и увлекательной задачи человечества.

Эта книга так и осталась бы пережитком принятия желаемого за действительное, если бы не мой соавтор Дана Маккензи, которая серьезно отнесся к моим желаниям и воплотил их в реальность. Он не только исправил мой иностранный акцент, но и повел меня в далекие страны, от кораблей военно-морского флота капитана Джеймса Линда до антарктической экспедиции капитана Роберта Скотта, добавив знания, истории, структуру и ясность к беспорядку математических уравнений, которые ожидали упорядочивающего повествования.

Я в большом долгу перед сотрудниками Лаборатории когнитивных систем Калифорнийского университета в Лос-Анджелесе, чьи работы и идеи за последние три с половиной десятилетия легли в научную основу этой книги: Алексом Балком, Элиасом Барейнбоймом, Блаем Бонетом, Карло Брито, Авином Ченом, Брайантом Ченом, Дэвидом Чикерингом, Аднаном Дарвичем, Риной Дехтер., Эндрю Форни, Дэвид Галлес, Гектор Геффнер, Дэн Гейгер, Мойзес Голдшмидт, Дэвид Хекерман, Марк Хопкинс, Джин Ким, Манабу Куроки, Трент Кионо, Картика Мохан, Азария Паз, Джордж Ребане, Илья Шпицер, Джин Тянь и Томас Верма. Финансирующие агентства получают ритуальную благодарность в научных публикациях, но слишком мало реального признания, учитывая их решающую роль в распознавании зародышей идей до того, как они станут модными. Я должен отметить постоянную и неизменную поддержку Национального научного фонда и Управления военно-морских исследований в рамках программы машинного обучения и разведки, возглавляемой Бехзадом Камгар-Парси.

Мы с Даной хотели бы поблагодарить нашего агента Джона Брокмана, который своевременно поддержал нас и воспользовался своим профессиональным опытом. Наш редактор *B*asic *B*ooks Ти Джей Келлехер задал нам правильные вопросы и убедил *B*asic *B*ooks в том, что такую амбициозную историю невозможно рассказать на 200 страницах. Наши иллюстраторы, Мааян Харел и Дакота Харр, сумели справиться с нашими иногда противоречивыми инструкциями и воплотили абстрактные

сюжеты в жизнь с юмором и красотой. Каору Малвихилл из Калифорнийского университета в Лос-Анджелесе заслуживает большой похвалы за проверку нескольких версий рукописи и иллюстрации множества графиков и диаграмм. Дана будет вечно благодарен Джону Уилксу, который основал Программу научных коммуникаций в Калифорнийском университете Санта-Крус, которая до сих пор набирает обороты и является наилучшим возможным путем к карьере научного писателя. Дана также хотел бы поблагодарить свою жену Кей, которая поощряла его следовать своей детской мечте стать писателем, даже когда это означало поднять ставки, пересечь страну и начать все сначала.

Наконец, я в глубочайшем долгу перед своей семьей за их терпение, понимание и поддержку. Особенно моей жене Рут, моему моральному компасу, за ее бесконечную любовь и мудрость. Моему покойному сыну Дэнни за то, что он показал мне молчаливую дерзость истины. Моим дочерям Тамаре и Мишель за то, что они поверили моему многолетнему обещанию, что книга в конце концов будет готова. И моим внукам, Леоре, Тори, Адаму, Ари и Эвану, за то, что они придавали смысл моим долгим путешествиям и всегда разрешали мои вопросы “почему”.

Заметки

ЗАМЕТКИ К ПРЕДИСЛОВИЮ

Students are never allowed: With possibly one exception: if we have performed a ran*do*mized controlled trial, as discussed in Chapter 4.

ЗАМЕТКИ К ГЛАВЕ ПЕРВОЙ

then the opposite is true: In other words, when evaluating an intervention in a causal model, we make the minimum changes possible to enforce its immediate effect. So we "break" the model where it comes to A but not B.

We should thank the language: I should also mention here that counterfactuals allow us to talk about causality in individual cases: What would have happened to Mr. Smith, who was not vaccinated and died of smallpox, if he had been vaccinated? Such questions, the backbone of personalized medicine, cannot be answered from rung-two information.

Yet we can answer: To be more precise, in geometry, undefined terms like "point" and "line" are primitives. The primitive in causal inference is the relation of "listening to," indicated by an arrow.

ЗАМЕТКИ К ГЛАВЕ ВТОРОЙ

And now the algebraic magic: For anyone who takes the trouble to read Wright's paper, let me warn you that he *do*es not compute his

path coefficients in grams per day. He computes them in "standard units" and then converts to grams per day at the end.

ЗАМЕТКИ К ГЛАВЕ ПЯТОЙ

"Cigarette smoking is causally related": The evidence for women was less clear at that time, primarily because women had smoked much less than men in the early decades of the century.

ЗАМЕТКИ К ГЛАВЕ ВОСЬМОЙ

And Abraham drew near: As before, I have used the King James translation but made small changes to align it more closely with the Hebrew.

The ease and familiarity of such: The 2013 Joint Statistical Meetings dedicated a whole session to the topic "Causal Inference as a Missing Data Problem"—Rubin's traditional mantra. One provocative paper at that session was titled "What Is Not a Missing Data Problem?" This title sums up my thoughts precisely.

This difference in commitment: Readers who are seeing this distinction for the first time should not feel alone; there are well over 100,000 regression analysts in the United States who are confused by this very issue, together with most authors of statistical textbooks. Things will only change when readers of this book take those authors to task.

Unfortunately, Rubin *do*es not consider: "Pearl's work is clearly interesting, and many researchers find his arguments that path diagrams are a natural and convenient way to express assumptions about causal structures appealing. In our own work, per- haps influenced by the type of examples arising in social and medical sciences, we have not found this approach to aid the drawing of causal inferences" (Imbens and Rubin 2013, p. 25).

One obstacle I faced was cyclic models: These are models with arrows that form a loop. I have avoided discussing them in this book, but such models are quite important in economics, for example.

Even today modern-day economists: Between 1995 and 1998, I presented the following toy puzzle to hundreds of econometrics stu-

dents and faculty across the United States:
Consider the classical supply-and-demand equations that every economics student solves in Economics 101.

1. What is the expected value of the demand Q if the price is reported to be P = p0?
2. What is the expected value of the demand Q if the price is set to P = p0?
3. Given that the current price is P = p0, what would the expected value of the demand Q be if we were to set the price at P = p1?

The reader should recognize these queries as coming from the three levels of the Ladder of Causation: predictions, actions, and counterfactuals. As I expected, respondents had no trouble answering question 1, one person (a distinguished professor) was able to solve question 2, and nobody managed to answer question 3.

The Model Penal Code expresses: This is a set of standard legal principles proposed by the American Law Institute in 1962 to bring uniformity to the various state legal codes. It *do*es not have full legal force in any state, but according to Wikipedia, as of 2016, more than two-thirds of the states have enacted parts of the Model Penal Code.

ЗАМЕТКИ К ГЛАВЕ ДЕВЯТОЙ

Those sailors who had eaten: The reason is that polar bear livers *do* contain vitamin C.

"On the Inadequacy of the Partial": The title refers to partial correlation, a standard method of controlling for a confounder that we discussed in Chapter 7.

Here is how to define the NIE: In the original delivery room, NIE was expressed using nested subscripts, as in $Y(0,M)$. I hope the reader will find the mixture of counterfactual subscripts and *do*-operators above more transparent.

In that year researchers identified: To be technically correct it should be called a "single nucleotide polymorphism," or SNP. It is a single letter in the genetic code, while a gene is more like a word or a sentence. However, in order not to burden the reader with unfamiliar terminology, I will simply refer to it as a gene.

Библиография

ВВЕДЕНИЕ: УМ ВАЖНЕЕ ДАННЫХ

Annotated Bibliography

The history of probability and statistics from antiquity to modern days is covered in depth by Hacking (1990); Stigler (1986, 1999, 2016). A less technical account is given in Salsburg (2002). Comprehensive accounts of the history of causal thought are unfortunately lacking, though interesting material can be found in Hoover (2008); Kleinberg (2015); Losee (2012); Mumford and Anjum (2014). The prohibition on causal talk can be seen in almost every standard statistical text, for example, Freedman, Pisani, and Purves (2007) or Efron and Hastie (2016). For an analysis of this prohibition as a linguistic impediment, see Pearl (2009, Chapters 5 and 11), and as a cultural barrier, see Pearl (2000b). Recent accounts of the achievements and limitations of Big Data and machine learning are Darwiche (2017); Pearl (2017); Mayer- Schönberger and Cukier (2013); *Do*mingos (2015); Marcus (July 30, 2017). Toulmin (1961) provides historical context to this debate. Readers interested in "model discovery" and more technical treatments of the *do*-operator can consult Pearl (1994, 2000a, Chapters 2–3); Spirtes, Glymour, and Scheines (2000). For a gentler introduc- tion, see Pearl, Glymour, and Jewell (2016). This last source is rec- ommended for readers with college-level mathematical skills but no background in statistics or computer science.

It also provides basic introduction to conditional probabilities, Bayes's rule, regression, and graphs.

Earlier versions of the inference engine shown in Figure 1.1 can be found in Pearl (2012); Pearl and Bareinboim (2014).

References

Darwiche, A. (2017). Human-level intelligence or animal-like abilities? Tech. rep., Department of Computer Science, University of California, Los Angeles, CA. Submitted to Communications of the ACM. Accessed online at https://arXiv:1707.04327.

*Do*mingos, P. (2015). The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World. Basic Books, New York, NY.

Efron, B., and Hastie, T. (2016). Computer Age Statistical Inference. Cambridge University Press, New York, NY.

Freedman, D., Pisani, R., and Purves, R. (2007). Statistics. 4th ed. W. W. Norton & Company, New York, NY.

Hacking, I. (1990). The Taming of Chance (Ideas in Context). Cambridge University Press, Cambridge, UK.

Hoover, K. (2008). Causality in economics and econometrics. In The New Palgrave Dictionary of Economics (S. Durlauf and L. Blume, eds.), 2nd ed. Palgrave Macmillan, New York, NY.

Kleinberg, S. (2015). Why: A Guide to Finding and Using Causes. O'Reilly Media, Sebastopol, CA.

Losee, J. (2012). Theories of Causality: From Antiquity to the Present. Routledge, New York, NY.

Marcus, G. (July 30, 2017). Artificial intelligence is stuck. Here's how to move it forward. New York Times, SR6.

Mayer-Schönberger, V., and Cukier, K. (2013). Big Data: A Revolution That Will Transform How We Live, Work, and Think. Houghton Mifflin Harcourt Publishing, New York, NY.

Morgan, S., and Winship, C. (2015). Counterfactuals and Causal Inference: Methods and Principles for Social Research (Analytical Methods for Social Research). 2nd ed. Cambridge University Press, New York, NY.

Mumford, S., and Anjum, R. L. (2014). Causation: A Very Short Introduction (Very Short Introductions). Oxford University Press, New York, NY.

Pearl, J. (1988). Probabilistic Reasoning in Intelligent Systems. Morgan Kaufmann, San Mateo, CA.

Pearl, J. (1994). A probabilistic calculus of actions. In Uncertainty in Artificial Intelligence 10 (R. L. de Mantaras and D. Poole, eds.). Morgan Kaufmann, San Mateo, CA, 454–462.

Pearl, J. (1995). Causal diagrams for empirical research. Biometrika 82: 669–710.

Pearl, J. (2000a). Causality: Models, Reasoning, and Inference. Cambridge University Press, New York, NY.

Pearl, J. (2000b). Comment on A. P. Dawid's Causal inference without counterfactuals. Journal of the American Statistical Association 95: 428–431.

Pearl, J. (2009). Causality: Models, Reasoning, and Inference. 2nd ed. Cambridge University Press, New York, NY.

Pearl, J. (2012). The causal foundations of structural equation modeling. In Handbook of Structural Equation Modeling (R. Hoyle, ed.). Guilford Press, New York, NY, 68–91.

Pearl, J. (2017). Advances in deep neural networks, at ACM Turing 50 Celebration. Available at: https://www.youtube.com/watch?v=mFYM9j8bGtg (June 23, 2017).

Pearl, J., and Bareinboim, E. (2014). External validity: From *do*-calculus to transportability across populations. Statistical Science 29:579–595.

Pearl, J., Glymour, M., and Jewell, N. (2016). Causal Inference in Statistics: A Primer. Wiley, New York, NY.

Provine, W. B. (1986). Sewall Wright and Evolutionary Biology. University of Chicago Press, Chicago, IL.

Salsburg, D. (2002). The Lady Tasting Tea: How Statistics Revolutionized Science in the Twentieth Century. Henry Holt and Company, LLC, New York, NY.

Spirtes, P., Glymour, C., and Scheines, R. (2000). Causation, Prediction, and Search. 2nd ed. MIT Press, Cambridge, MA.

Stigler, S. M. (1986). The History of Statistics: The Measurement of Uncertainty Before 1900. Belknap Press of Harvard University Press, Cambridge, MA.

Stigler, S. M. (1999). Statistics on the Table: The History of Statistical Concepts and Methods. Harvard University Press, Cambridge, MA.

Stigler, S. M. (2016). The Seven Pillars of Statistical Wis*do*m. Harvard University Press, Cambridge, MA.

Toulmin, S. (1961). Foresight and Understanding: An Enquiry into the Aims of Science. University of Indiana Press, Bloomington, IN. Virgil. (29 bc). Georgics. Verse 490, Book 2.

ГЛАВА 1. ЛЕСТНИЦА ПРИЧИННОСТИ

Annotated Bibliography

A technical account of the distinctions between the three levels of the Ladder of Causation can be found in Chapter 1 of Pearl (2000).

Our comparisons between the Ladder of Causation and human cognitive development were inspired by Harari (2015) and by the recent findings by Kind et al. (2014). Kind's article contains details about the Lion Man and the site where it was found. Related research on the development of causal understanding in babies can be found in Weisberg and Gopnik (2013).

The Turing test was first proposed as an imitation game in 1950 (Turing, 1950). Searle's "Chinese Room" argument appeared in Searle (1980) and has been widely discussed in the years since. See Russell and Norvig (2003); Preston and Bishop (2002); Pinker (1997).

The use of model modification to represent intervention has its conceptual roots with the economist Trygve Haavelmo (1943); see Pearl (2015) for a detailed account. Spirtes, Glymour, and Scheines (1993) gave it a graphical representation in terms of arrow deletion. Balke and Pearl (1994a, 1994b) extended it to simulate counterfactual reasoning, as demonstrated in the firing squad example.

A comprehensive summary of probabilistic causality is given in Hitchcock (2016). Key ideas can be found in Reichenbach (1956); Sup- pes (1970); Cartwright (1983); Spohn (2012). My analyses of proba- bilistic causality and probability raising are presented in Pearl (2000; 2009, Section 7.5; 2011).

References

Balke, A., and Pearl, J. (1994a). Counterfactual probabilities: Computational methods, bounds, and applications. In Uncertainty in Artificial Intelligence 10 (R. L. de Mantaras and D. Poole, eds.). Morgan Kaufmann, San Mateo, CA, 46–54.

Balke, A., and Pearl, J. (1994b). Probabilistic evaluation of counterfactual queries. In Proceedings of the Twelfth National Conference on Artificial Intelligence, vol. 1. MIT Press, Menlo Park, CA, 230–237.

Cartwright, N. (1983). How the Laws of Physics Lie. Clarendon Press, Oxford, UK.

Haavelmo, T. (1943). The statistical implications of a system of simultaneous equations. Econometrica 11: 1–12. Reprinted in D. F. Hendry and M. S. Morgan (Eds.), The Foundations of Econometric Analysis, Cambridge University Press, Cambridge, UK, 477–490, 1995.

Harari, Y. N. (2015). Sapiens: A Brief History of Humankind. Harper Collins Publishers, New York, NY.

Hitchcock, C. (2016). Probabilistic causation. In Stanford Encyclopedia of Philosophy (Winter 2016) (E. N. Zalta, ed.). Metaphysics Research Lab, Stanford, CA. Available at: https://stanford.library.sydney.edu.au/archives/win2016/entries/causation-probabilistic. Kind, C.-J., Ebinger-Rist, N., Wolf, S., Beutelspacher, T., and Wehr- berger, K. (2014). The smile of the Lion Man. Recent excavations

in Stadel cave (Baden-Württemberg, south-western Germany) and the restoration of the famous upper palaeolithic figurine. Quartär 61: 129–145.

Pearl, J. (2000). Causality: Models, Reasoning, and Inference. Cambridge University Press, New York, NY.

Pearl, J. (2009). Causality: Models, Reasoning, and Inference. 2nd ed. Cambridge University Press, New York, NY.

Pearl, J. (2011). The structural theory of causation. In Causality in the Sciences (P. M. Illari, F. Russo, and J. Williamson, eds.), chap. 33. Clarendon Press, Oxford, UK, 697–727.

Pearl, J. (2015). Trygve Haavelmo and the emergence of causal calculus. Econometric Theory 31: 152–179. Special issue on Haavelmo centennial.

Pinker, S. (1997). How the Mind Works. W. W. Norton and Company, New York, NY.

Preston, J., and Bishop, M. (2002). Views into the Chinese Room: New Essays on Searle and Artificial Intelligence. Oxford University Press, New York, NY.

Reichenbach, H. (1956). The Direction of Time. University of California Press, Berkeley, CA.

Russell, S. J., and Norvig, P. (2003). Artificial Intelligence: A Modern Approach. 2nd ed. Prentice Hall, Upper Saddle River, NJ.

Searle, J. (1980). Minds, brains, and programs. Behavioral and Brain Sciences 3: 417–457.

Spirtes, P., Glymour, C., and Scheines, R. (1993). Causation, Prediction, and Search. Springer-Verlag, New York, NY.

Spohn, W. (2012). The Laws of Belief: Ranking Theory and Its Philosophical Applications. Oxford University Press, Oxford, UK.

Suppes, P. (1970). A Probabilistic Theory of Causality. North-Holland Publishing Co., Amsterdam, Netherlands.

Turing, A. (1950). Computing machinery and intelligence. Mind 59: 433–460.

Weisberg, D. S., and Gopnik, A. (2013). Pretense, counterfactuals, and Bayesian causal models: Why what is not real really matters. Cognitive Science 37: 1368–1381.

ГЛАВА 2. ОТ ГОСУДАРСТВЕННЫХ ПИРАТОВ ДО МОРСКИХ СВИНОК: СТАНОВЛЕНИЕ ПРИЧИННОГО ВЫВОДА

Annotated Bibliography

Galton's explorations of heredity and correlation are described in his books (Galton, 1869, 1883, 1889) and are also *do*cumented in Stigler (2012, 2016).

For a basic introduction to the Hardy-Weinberg equilibrium, see Wikipedia (2016a). For the origin of Galileo's quote "E pur si muove," see Wikipedia (2016b). The story of the Paris catacombs and Pearson's shock at correlations induced by "artificial mixtures" can be found in Stigler (2012, p. 9).

Because Wright lived such a long life, he had the rare privilege of seeing a biography (Provine, 1986) come out while he was still alive. Provine's biography is still the best place to learn about Wright's career, and we particularly recommend Chapter 5 on path analysis. Crow's two biographical sketches (Crow, 1982, 1990) also provide a very useful biographical perspective. Wright (1920) is the seminal paper on path diagrams; Wright (1921) is a fuller exposition and

the source for the guinea pig birth-weight example. Wright (1983) is Wright's response to Karlin's critique, written when he was over ninety years old.

The fate of path analysis in economics and social science is narrated in Chapter 5 of Pearl (2000) and in Bollen and Pearl (2013). Blalock (1964), Duncan (1966), and Goldberger (1972) introduced Wright's ideas to social science with great enthusiasm, but their theoretical underpinnings were not well articulated. A decade later, when Freedman (1987) challenged path analysts to explain how interventions are modelled, the enthusiasm disappeared, and leading researchers retreated to viewing SEM as an exercise in statistical analysis. This revealing discussion among twelve scholars is *do*cumented in the same issue of the Journal of Educational Statistics as Freedman's article.

The reluctance of economists to embrace diagrams and structural notation is described in Pearl (2015). The painful consequences for economic education are *do*cumented in Chen and Pearl (2013).

A popular exposition of the Bayesian-versus-frequentist debate is given in McGrayne (2011).

More technical discussions can be found in Efron (2013) and Lind- ley (1987).

References

Blalock, H., Jr. (1964). Causal Inferences in Nonexperimental Research. University of North Carolina Press, Chapel Hill, NC.

Bollen, K., and Pearl, J. (2013). Eight myths about causality and structural equation models. In Handbook of Causal Analysis for Social Research (S. Morgan, ed.). Springer, *Do*rdrecht, Netherlands, 301–328.

Chen, B., and Pearl, J. (2013). Regression and causation: A critical examination of econometrics textbooks. Real-World Economics Review 65: 2–20.

Crow, J. F. (1982). Sewall Wright, the scientist and the man. Perspectives in Biology and Medicine 25: 279–294.

Crow, J. F. (1990). Sewall Wright's place in twentieth-century biology. Journal of the History of Biology 23: 57–89.

Duncan, O. D. (1966). Path analysis. American Journal of Sociology 72: 1–16.

Efron, B. (2013). Bayes' theorem in the 21st century. Science 340: 1177–1178.

Freedman, D. (1987). As others see us: A case study in path analysis (with discussion). Journal of Educational Statistics 12: 101–223. Galton, F. (1869). Hereditary Genius. Macmillan, Lon*d*on, UK. Galton, F. (1883). Inquiries into Human Faculty and Its Development.

Macmillan, Lon*d*on, UK.

Galton, F. (1889). Natural Inheritance. Macmillan, Lon*d*on, UK.

Goldberger, A. (1972). Structural equation models in the social sciences. Econometrica: Journal of the Econometric Society 40: 979–1001.

Lindley, D. (1987). Bayesian Statistics: A Review. CBMS-NSF Regional Conference Series in Applied Mathematics (Book 2). Society for Industrial and Applied Mathematics, Philadelphia, PA.

McGrayne, S. B. (2011). The Theory That Would Not Die. Yale University Press, New Haven, CT.

Pearl, J. (2000). Causality: Models, Reasoning, and Inference. Cambridge University Press, New York, NY.

Pearl, J. (2015). Trygve Haavelmo and the emergence of causal calculus. Econometric Theory 31: 152–179. Special issue on Haavelmo centennial.

Provine, W. B. (1986). Sewall Wright and Evolutionary Biology. University of Chicago Press, Chicago, IL.

Stigler, S. M. (2012). Studies in the history of probability and statistics, L: Karl Pearson and the rule of three. Biometrika 99: 1–14.

Stigler, S. M. (2016). The Seven Pillars of Statistical Wis*d*om. Harvard University Press, Cambridge, MA.

Wikipedia. (2016a). Hardy-Weinberg principle. Available at: https://en.wikipedia.org/wiki/Hardy-Weinberg-principle (last edited: Oc- tober 2, 2016).

Wikipedia. (2016b). Galileo Galilei. Available at: https://en.wikipedia.org/wiki/Galileo_Galilei (last edited: October 6, 2017).

Wright, S. (1920). The relative importance of heredity and environment in determining the piebald pattern of guinea-pigs. Proceedings of the National Academy of Sciences of the United States of America 6: 320–332.

Wright, S. (1921). Correlation and causation. Journal of Agricultural Research 20: 557–585.
Wright, S. (1983). On "Path analysis in genetic epidemiology: A critique." American Journal of Human Genetics 35: 757–768.

ГЛАВА 3. ОТ ДОКАЗАТЕЛЬСТВ К ПРИЧИНАМ. ПРЕПОДОБНЫЙ БАЙЕС ЗНАКОМИТСЯ С МИСТЕРОМ ХОЛМСОМ

Annotated Bibliography

Elementary introductions to Bayes's rule and Bayesian thinking can be found in Lindley (2014) and Pearl, Glymour, and Jewell (2016).

Debates with competing representations of uncertainty are presented in Pearl (1988); see also the extensive list of references given there.

Our mammogram data are based primarily on information from the Breast Cancer Surveillance Consortium (BCSC, 2009) and US Preventive Services Task Force (USPSTF, 2016) and are presented for instructional purposes only.

"Bayesian networks" received their name in 1985 (Pearl, 1985) and were first presented as a model of self-activated memory. Applications to expert systems followed the development of belief updating algorithms for loopy networks (Pearl, 1986; Lauritzen and Spiegelhalter, 1988).

The concept of d-separation, which connects path blocking in a diagram to dependencies in the data, has its roots in the theory of graphoids (Pearl and Paz, 1985). The theory unveils the common properties of graphs (hence the name) and probabilities and explains why these two seemingly alien mathematical objects can support one another in so many ways. See also "Graphoid," Wikipedia.

The amusing example of the bag on the airline flight can be found in Conrady and Jouffe (2015, Chapter 4).

The Malaysia Airlines Flight 17 disaster was well covered in the media; see Clark and Kramer (October 14, 2015) for an update on the investigation a year after the incident. Wiegerinck, Burgers, and Kappen (2013) describes how Bonaparte works. Further details on the identification of Flight 17 victims, including the pedigree shown

in Figure 3.7, came from personal correspondence from W. Burgers to D. Mackenzie (August 24, 2016) and from a phone interview with W. Burgers and B. Kappen by D. Mackenzie (August 23, 2016).

The complex and fascinating story of turbo and low-density parity-check codes has not been told in a truly layman-friendly form, but good starting points are Costello and Forney (2007) and Hardesty (2010a, 2010b). The crucial realization that turbo codes work by the belief propagation algorithm stems from McEliece, David, and Cheng (1998). Efficient codes continue to be a battleground for wireless communications; Carlton (2016) takes a look at the current contenders for "5G" phones (due out in the 2020s).

References

Breast Cancer Surveillance Consortium (BCSC). (2009). Performance measures for 1,838,372 screening mammography examinations from 2004 to 2008 by age. Available at: http://www.bcsc-research .org/statistics/performance/screening/2009/perf_age.html (accessed October 12, 2016).

Carlton, A. (2016). Surprise! Polar codes are coming in from the cold. Computerworld. Available at: https://www.computerworld.com/article/3151866/mobile-wireless/surprise-polar-codes-are-coming-in-from-the-cold.html (posted December 22, 2016).

Clark, N., and Kramer, A. (October 14, 2015). Malaysia Airlines Flight 17 most likely hit by Russian-made missile, inquiry says. New York Times.

Conrady, S., and Jouffe, L. (2015). Bayesian Networks and Bayesia Lab: A Practical Introduction for Researchers. Bayesia USA, Franklin, TN.

Costello, D. J., and Forney, G. D., Jr. (2007). Channel coding: The road to channel capacity. Proceedings of IEEE 95: 1150–1177. Hardesty, L. (2010a). Explained: Gallager codes. MIT News. Available at: http://news.mit.edu/2010/gallager-codes-0121 (posted: January 21, 2010).

Hardesty, L. (2010b). Explained: The Shannon limit. MIT News. Available at: http://news.mit.edu/2010/explained-shannon-0115 (posted January 19, 2010).

Lauritzen, S., and Spiegelhalter, D. (1988). Local computations with probabilities on graphical structures and their application to expert systems (with discussion). Journal of the Royal Statistical Society, Series B 50: 157–224.

Lindley, D. V. (2014). Understanding Uncertainty. Rev. ed. John Wiley and Sons, Hoboken, NJ.

McEliece, R. J., David, J. M., and Cheng, J. (1998). Turbo decoding as an instance of Pearl's "belief propagation" algorithm. IEEE Journal on Selected Areas in Communications 16: 140–152.

Pearl, J. (1985). Bayesian networks: A model of self-activated memory for evidential reasoning. In Proceedings, Cognitive Science Society (CSS-7). UCLA Computer Science Department, Irvine, CA.

Pearl, J. (1986). Fusion, propagation, and structuring in belief networks. Artificial Intelligence 29: 241–288.

Pearl, J. (1988). Probabilistic Reasoning in Intelligent Systems. Morgan Kaufmann, San Mateo, CA.

Pearl, J., Glymour, M., and Jewell, N. (2016). Causal Inference in Statistics: A Primer. Wiley, New York, NY.

Pearl, J., and Paz, A. (1985). GRAPHOIDS: A graph-based logic for reasoning about relevance relations. Tech. Rep. 850038 (R-53-L). Computer Science Department, University of California, Los An- geles. Short version in B. DuBoulay, D. Hogg, and L. Steels (Eds.) Advances in Artificial Intelligence—II, Amsterdam, North Holland, 357–363, 1987.

US Preventive Services Task Force (USPSTF) (2016). Final recommendation statement: Breast cancer: Screening. Available at: https://www.uspreventiveservicestaskforce.org/Page/*Do*cument/RecommendationStatementFinal/breast-cancer-screening1 (up- dated: January 2016).

Wikipedia. (2018). Graphoid. Available at: https://en.wikipedia.org/wiki/Graphoid (last edited: January 8, 2018).

Wiegerinck, W., Burgers, W., and Kappen, B. (2013). Bayesian networks, introduction and practical applications. In Handbook on Neural Information Processing (M. Bianchini, M. Maggini, and L. C. Jain, eds.). Intelligent Systems Reference Library (Book 49). Springer, Berlin, Germany, 401–431.

ГЛАВА 4. ОСЛОЖНИТЕЛИ И НАОБОРОТ: КАК УБИТЬ ПРЯЧУЩУЮСЯ ПЕРЕМЕННУЮ

Annotated Bibliography

The story of Daniel has frequently been cited as the first controlled trial; see, for example, Lilienfeld (1982) or Stigler (2016). The results of the Honolulu walking study were reported in Hakim (1998).

Fisher Box's lengthy quote about "the skillful interrogation of Nature" comes from her excellent biography of her father (Box, 1978, Chapter 6). Fisher, too, wrote about experiments as a dialogue with Nature; see Stigler (2016). Thus I believe we can think of her quote as nearly coming from the patriarch himself, only more beautifully expressed.

It is fascinating to read Weinberg's papers on confounding (Weinberg, 1993; Howards et al., 2012) back-to-back. They are like two snapshots of the history of confounding, one taken just before causal diagrams became widespread and the second taken twenty years later, revisiting the same examples using causal diagrams. Forbes's complicated diagram of the causal network for asthma and smoking can be found in Williamson et al. (2014).

Morabia's "classic epidemiological definition of confounding" can be found in Morabia (2011). The quotes from David Cox come from Cox (1992, pp. 66–67). Other good sources on the history of confounding are Greenland and Robins (2009) and Wikipedia (2016).

The back-*do*or criterion for eliminating confounding bias, together with its adjustment formula, were introduced in Pearl (1993). Its impact on epidemiology can be seen through Greenland, Pearl, and Robins (1999). Extensions to sequential interventions and other nuances are developed in Pearl (2000, 2009) and more gently described in Pearl, Glymour, and Jewell (2016). Software for computing causal effects using *do*-calculus is available in Tikka and Karvanen (2017).

The paper by Greenland and Robins (1986) was revisited by the authors a quarter century later, in light of the extensive developments since that time, including the advent of causal diagrams (Greenland and Robins, 2009).

References

Box, J. F. (1978). R. A. Fisher: The Life of a Scientist. John Wiley and Sons, New York, NY.

Cox, D. (1992). Planning of Experiments. Wiley-Interscience, New York, NY.

Greenland, S., Pearl, J., and Robins, J. (1999). Causal diagrams for epidemiologic research. Epidemiology 10: 37–48.

Greenland, S., and Robins, J. (1986). Identifiability, exchangeability, and epidemiological confounding. International Journal of Epidemiology 15: 413–419.

Greenland, S., and Robins, J. (2009). Identifiability, exchangeability, and confounding revisited. Epidemiologic Perspectives & Innovations 6. *do*i:10.1186/1742-5573-6-4.

Hakim, A. (1998). Effects of walking on mortality among nonsmoking retired men. New England Journal of Medicine 338: 94–99.

Hernberg, S. (1996). Significance testing of potential confounders and other properties of study groups—Misuse of statistics. Scandinavian Journal of Work, Environment and Health 22: 315–316.

Howards, P. P., Schisterman, E. F., Poole, C., Kaufman, J. S., and Weinberg, C. R. (2012). "Toward a clearer definition of confounding" revisited with directed acyclic graphs. American Journal of Epidemiology 176: 506–511.

Lilienfeld, A. (1982). Ceteris paribus: The evolution of the clinical trial. Bulletin of the History of Medicine 56: 1–18.

Morabia, A. (2011). History of the modern epidemiological concept of confounding. Journal of Epidemiology and Community Health 65: 297–300.

Pearl, J. (1993). Comment: Graphical models, causality, and interven- tion. Statistical Science 8: 266–269.

Pearl, J. (2000). Causality: Models, Reasoning, and Inference. Cambridge University Press, New York, NY.

Pearl, J. (2009). Causality: Models, Reasoning, and Inference. 2nd ed. Cambridge University Press, New York, NY.

Pearl, J., Glymour, M., and Jewell, N. (2016). Causal Inference in Statistics: A Primer. Wiley, New York, NY.

Stigler, S. M. (2016). The Seven Pillars of Statistical Wis*do*m. Harvard University Press, Cambridge, MA.

Tikka, J., and Karvanen, J. (2017). Identifying causal effects with the R Package causaleffect. Journal of Statistical Software 76, no. 12. *do*i:10.18637/jss.r076.i12.

Weinberg, C. (1993). Toward a clearer definition of confounding.

American Journal of Epidemiology 137: 1–8.

Wikipedia. (2016). Confounding. Available at: https://en.wikipedia.org/wiki/Confounding (accessed: September 16, 2016). Williamson, E., Aitken, Z., Lawrie, J., Dharmage, S., Burgess, H., and Forbes, A. (2014). Introduction to causal diagrams for confounder selection. Respirology 19: 303–311.

ГЛАВА 5. ДЫМНЫЕ ДЕБАТЫ: НА СВЕЖИЙ ВОЗДУХ

Annotated Bibliography

Two book-length studies, Brandt (2007) and Proctor (2012a), contain all the information any reader could ask for about the smoking–lung cancer debate, short of reading the actual tobacco company *do*cuments (which are available online). Shorter surveys of the smoking-cancer debate in the 1950s are Salsburg (2002, Chapter 18), Parascan*do*la (2004), and Proctor (2012b). Stolley (1991) takes a look at the unique role of R. A. Fisher, and Greenhouse (2009) comments on Jerome Cornfield's importance. The shot heard around the world was *Do*ll and Hill (1950), which first implicated smoking in lung cancer; though technical, it is a scientific classic.

For the story of the surgeon general's committee and the emergence of the Hill guidelines for causation, see Blackburn and Labarthe (2012) and Morabia (2013). Hill's own description of his criteria can be found in Hill (1965).

Lilienfeld (2007) is the source of the "Abe and Yak" story with which we began the chapter.

VanderWeele (2014) and Hernández-Díaz, Schisterman, and Hernán (2006) resolve the birth-weight para*do*x using causal diagrams. An interesting "before-and-after" pair of articles is Wilcox (2001, 2006), written before and after the author learned about causal diagrams; his excitement in the latter article is palpable.

Readers interested in the latest statistics and historical trends in cancer mortality and smoking may consult US Department of Health and Human Services (USDHHS, 2014), American Cancer Society (2017), and Wingo (2003).

References

American Cancer Society. (2017). Cancer facts and figures. Available at: https://www.cancer.org/research/cancer-facts-statistics.html (posted: February 19, 2015).

Blackburn, H., and Labarthe, D. (2012). Stories from the evolution of guidelines for causal inference in epidemiologic associations: 1953–1965. American Journal of Epidemiology 176: 1071–1077. Brandt, A. (2007). The Cigarette Century. Basic Books, New York, NY.

*Do*ll, R., and Hill, A. B. (1950). Smoking and carcinoma of the lung. British Medical Journal 2: 739–748.

Greenhouse, J. (2009). Commentary: Cornfield, epidemiology, and causality. International Journal of Epidemiology 38: 1199–1201.

Hernández-Díaz, S., Schisterman, E., and Hernán, M. (2006). The birth weight "para*do*x" uncovered? American Journal of Epidemiology 164: 1115–1120.

Hill, A. B. (1965). The environment and disease: Association or causation? Journal of the Royal Society of Medicine 58: 295–300.

Lilienfeld, A. (2007). Abe and Yak: The interactions of Abraham M.

Lilienfeld and Jacob Yerushalmy in the development of modern epidemiology (1945–1973). Epidemiology 18: 507–514.

Morabia, A. (2013). Hume, Mill, Hill, and the sui generis epidemiologic approach to causal inference. American Journal of Epidemiology 178: 1526–1532.

Parascan*do*la, M. (2004). Two approaches to etiology: The debate over smoking and lung cancer in the 1950s. Endeavour 28: 81–86. Proctor, R. (2012a). Golden Holocaust: Origins of the Cigarette Catastrophe and the Case for Abolition. University of California Press, Berkeley, CA.

Proctor, R. (2012b). The history of the discovery of the cigarette–lung cancer link: Evidentiary traditions, corporate denial, and global toll. Tobacco Control 21: 87–91.

Salsburg, D. (2002). The Lady Tasting Tea: How Statistics Revolutionized Science in the Twentieth Century. Henry Holt and Company, New York, NY.

Stolley, P. (1991). When genius errs: R. A. Fisher and the lung cancer controversy. American Journal of Epidemiology 133: 416–425.

US Department of Health and Human Services (USDHHS). (2014).

The health consequences of smoking—50 years of progress: A report of the surgeon general. USDHHS and Centers for Disease Control and Prevention, Atlanta, GA.

VanderWeele, T. (2014). Commentary: Resolutions of the birthweight para*do*x: Competing explanations and analytical insights. Inter- national Journal of Epidemiology 43: 1368–1373.

Wilcox, A. (2001). On the importance—and the unimportance—of birthweight. International Journal of Epidemiology 30: 1233–1241. Wilcox, A. (2006). The perils of birth weight—A lesson from directed acyclic graphs. American Journal of Epidemiology 164: 1121–1123. Wingo, P. (2003). Long-term trends in cancer mortality in the United States, 1930–1998. Cancer 97: 3133–3275.

ГЛАВА 6: СПЛОШНЫЕ ПАРАДОКСЫ!

Annotated Bibliography

The Monty Hall para*do*x appears in many introductory books on probability theory (e.g., Grinstead and Snell, 1998, p. 136; Lindley, 2014, p. 201). The equivalent "three prisoners dilemma" was used to demonstrate the inadequacy of non-Bayesian approaches in Pearl (1988, pp. 58–62).

Tierney (July 21, 1991) and Crockett (2015) tell the amazing story of vos Savant's column on the Monty Hall para*do*x; Crockett gives several other entertaining and embarrassing comments that vos Savant received from so-called experts. Tierney's article tells what Monty Hall himself thought of the fuss—an interesting human-interest angle! An extensive account of the history of Simpson's para*do*x is given in Pearl (2009, pp. 174–182), including many attempts by statisticians and philosophers to resolve it without invoking causation. A more recent account, geared for educators, is given in Pearl (2014).

Savage (2009), Julious and Mullee (1994), and Appleton, French,

and Vanderpump (1996) give the three real-world examples of Simpson's para*do*x mentioned in the text (relating to baseball, kidney stones, and smoking, respectively).

Savage's sure-thing principle (Savage, 1954) is treated in Pearl (2016b), and its corrected causal version is derived in Pearl (2009, pp. 181–182).

Versions of Lord's para*do*x (Lord, 1967) are described in Glymour (2006); Hernández-Díaz, Schisterman, and Hernán (2006); Senn (2006); Wainer (1991). A comprehensive analysis can be found in Pearl (2016a).

Para*do*xes invoking counterfactuals are not included in this chapter but are no less intriguing. For a sample, see Pearl (2013).

References

Appleton, D., French, J., and Vanderpump, M. (1996). Ignoring a covariate: An example of Simpson's para*do*x. American Statistician 50: 340–341.

Crockett, Z. (2015). The time everyone "corrected" the world's smartest woman. Priceonomics. Available at: http://priceonomics.com/the-time-everyone-corrected-the-worlds-smartest (posted: February 19, 2015).

Glymour, M. M. (2006). Using causal diagrams to understand common problems in social epidemiology. In Methods in Social Epidemiology. John Wiley and Sons, San Francisco, CA, 393–428.

Grinstead, C. M., and Snell, J. L. (1998). Introduction to Probability.

2nd rev. ed. American Mathematical Society, Providence, RI. Hernández-Díaz, S., Schisterman, E., and Hernán, M. (2006). The

birth weight "para*do*x" uncovered? American Journal of Epide- miology 164: 1115–1120.

Julious, S., and Mullee, M. (1994). Confounding and Simpson's para*do*x. British Medical Journal 309: 1480–1481.

Lindley, D. V. (2014). Understanding Uncertainty. Rev. ed. John Wiley and Sons, Hoboken, NJ.

Lord, F. M. (1967). A para*do*x in the interpretation of group comparisons. Psychological Bulletin 68: 304–305.

Pearl, J. (1988). Probabilistic Reasoning in Intelligent Systems. Morgan Kaufmann, San Mateo, CA.

Pearl, J. (2009). Causality: Models, Reasoning, and Inference. 2nd ed. Cambridge University Press, New York, NY.

Pearl, J. (2013). The curse of free-will and para*do*x of inevitable regret. Journal of Causal Inference 1: 255–257.

Pearl, J. (2014). Understanding Simpson's para*do*x. American Statistician 88: 8–13.

Pearl, J. (2016a). Lord's para*do*x revisited—(Oh Lord! Kumbaya!). Journal of Causal Inference 4. *do*i:10.1515/jci-2016-0021.

Pearl, J. (2016b). The sure-thing principle. Journal of Causal Inference 4: 81–86.

Savage, L. (1954). The Foundations of Statistics. John Wiley and Sons, New York, NY.

Savage, S. (2009). The Flaw of Averages: Why We Underestimate Risk in the Face of Uncertainty. John Wiley and Sons, Hoboken, NJ.

Senn, S. (2006). Change from baseline and analysis of covariance revisited. Statistics in Medicine 25: 4334–4344.

Simon, H. (1954). Spurious correlation: A causal interpretation. Journal of the American Statistical Association 49: 467–479.

Tierney, J. (July 21, 1991). Behind Monty Hall's *do*ors: Puzzle, debate and answer? New York Times.

Wainer, H. (1991). Adjusting for differential base rates: Lord's para*do*x again. Psychological Bulletin 109: 147–151.

ГЛАВА 7: ЗА ПРЕДЕЛАМИ ПОПРАВОК: ПОКОРЕНИЕ ГОРЫ ИНТЕРВЕНЦИИ

Annotated Bibliography

Extensions of the back-*do*or and front-*do*or adjustments were first reported in Tian and Pearl (2002) based on Tian's c-component factorization. These were followed by Shpitser's algorithmization of the *do*-calculus (Shpitser and Pearl, 2006a) and then the completeness results of Shpitser and Pearl (2006b) and Huang and Valtorta (2006).

The economists among our readers should note that the cultural resistance of some economists to graphical tools of analysis (Heckman and Pinto, 2015; Imbens and Rubin, 2015) is not shared by all

economists. White and Chalak (2009), for example, have generalized and applied the *do*-calculus to economic systems involving equilibrium and learning. Recent textbooks in the social and behavioral sciences, Morgan and Winship (2007) and Kline (2016), further signal to young researchers that cultural ortho*do*xy, like the fear of telescopes in the seventeenth century, is not long lasting in the sciences.

John Snow's investigation of cholera was very little appreciated during his lifetime, and his one-paragraph obituary in Lancet did not even mention it. Remarkably, the premier British medical journal "corrected" its obituary 155 years later (Hempel, 2013). For more biographical material on Snow, see Hill (1955) and Cameron and Jones (1983). Glynn and Kashin (2018) is one of the first papers to demon- strate empirically that front-*do*or adjustment is superior to back-*do*or adjustment when there are unobserved confounders. Freedman's critique of the smoking–tar–lung cancer example can be found in a chapter of Freedman (2010) titled "On Specifying Graphical Models for Causation."

Introductions to instrumental variables can be found in Greenland (2000) and in many textbooks of econometrics (e.g., Bowden and Turkington, 1984; Wooldridge, 2013).

Generalized instrumental variables, extending the classical definition given in our text, were introduced in Brito and Pearl (2002).

The program DAGitty (available online at http://www.dagitty.net/dags.html) permits users to search the diagram for generalized instru- mental variables and reports the resulting estimands (Textor, Hardt, and Knüppel, 2011). Another diagram-based software package for decision making is BayesiaLab (www.bayesia.com).

Bounds on instrumental variable estimates are studied at length in Chapter 8 of Pearl (2009) and are applied to the problem of noncompliance. The LATE approximation is advocated and debated in Imbens (2010).

References

Bareinboim, E., and Pearl, J. (2012). Causal inference by surrogate experiments: z-identifiability. In Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence (N. de Freitas and K. Murphy, eds.). AUAI Press, Corvallis, OR.

Bowden, R., and Turkington, D. (1984). Instrumental Variables. Cambridge University Press, Cambridge, UK.

Brito, C., and Pearl, J. (2002). Generalized instrumental variables. In Uncertainty in Artificial Intelligence, Proceedings of the Eighteenth Conference (A. Darwiche and N. Friedman, eds.). Morgan Kaufmann, San Francisco, CA, 85–93.

Cameron, D., and Jones, I. (1983). John Snow, the Broad Street pump, and modern epidemiology. International Journal of Epidemiology 12: 393–396.

Cox, D., and Wermuth, N. (2015). Design and interpretation of stud- ies: Relevant concepts from the past and some extensions. Observational Studies 1. Available at: https://arxiv.org/pdf/1505.02452.pdf.

Freedman, D. (2010). Statistical Models and Causal Inference: A Dialogue with the Social Sciences. Cambridge University Press, New York, NY.

Glynn, A., and Kashin, K. (2018). Front-*do*or versus back-*do*or adjustment with unmeasured confounding: Bias formulas for front-*do*or and hybrid adjustments. Journal of the American Statistical Association. To appear.

Greenland, S. (2000). An introduction to instrumental variables for epidemiologists. International Journal of Epidemiology 29: 722–729. Heckman, J. J., and Pinto, R. (2015). Causal analysis after Haavelmo. Econometric Theory 31: 115–151.

Hempel, S. (2013). Obituary: John Snow. Lancet 381: 1269–1270.

Hill, A. B. (1955). Snow—An appreciation. Journal of Economic Perspectives 48: 1008–1012.

Huang, Y., and Valtorta, M. (2006). Pearl's calculus of intervention is complete. In Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence (R. Dechter and T. Richardson, eds.). AUAI Press, Corvallis, OR, 217–224.

Imbens, G. W. (2010). Better LATE than nothing: Some comments on Deaton (2009) and Heckman and Urzua (2009). Journal of Economic Literature 48: 399–423.

Imbens, G. W., and Rubin, D. B. (2015). Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction. Cambridge University Press, Cambridge, MA.

Kline, R. B. (2016). Principles and Practice of Structural Equation Modeling. 3rd ed. Guilford, New York, NY.

Morgan, S., and Winship, C. (2007). Counterfactuals and Causal Inference: Methods and Principles for Social Research (Analytical Methods for Social Research). Cambridge University Press, New York, NY.

Pearl, J. (2009). Causality: Models, Reasoning, and Inference. 2nd ed. Cambridge University Press, New York, NY.

Pearl, J. (2013). Reflections on Heckman and Pinto's "Causal analysis after Haavelmo." Tech. Rep. R-420. Department of Computer Science, University of California, Los Angeles, CA. Working paper.

Pearl, J. (2015). Indirect confounding and causal calculus (on three papers by Cox and Wermuth). Tech. Rep. R-457. Department of Computer Science, University of California, Los Angeles, CA.

Shpitser, I., and Pearl, J. (2006a). Identification of conditional interventional distributions. In Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence (R. Dechter and T. Richardson, eds.). AUAI Press, Corvallis, OR, 437–444.

Shpitser, I., and Pearl, J. (2006b). Identification of joint interventional distributions in recursive semi-Markovian causal models. In Proceedings of the Twenty-First National Conference on Artificial Intelligence. AAAI Press, Menlo Park, CA, 1219–1226.

Stock, J., and Trebbi, F. (2003). Who invented instrumental variable regression? Journal of Economic Perspectives 17: 177–194.

Textor, J., Hardt, J., and Knüppel, S. (2011). DAGitty: A graphical tool for analyzing causal diagrams. Epidemiology 22: 745.

Tian, J., and Pearl, J. (2002). A general identification condition for causal effects. In Proceedings of the Eighteenth National Conference on Artificial Intelligence. AAAI Press/MIT Press, Menlo Park, CA, 567–573.

Wermuth, N., and Cox, D. (2008). Distortion of effects caused by indirect confounding. Biometrika 95: 17–33. (See Pearl [2009, Chapter 4] for a general solution.)

Wermuth, N., and Cox, D. (2014). Graphical Markov models: Overview. ArXiv: 1407.7783.

White, H., and Chalak, K. (2009). Settable systems: An extension of Pearl's causal model with optimization, equilibrium and learning. Journal of Machine Learning Research 10: 1759–1799.

Wooldridge, J. (2013). Introductory Econometrics: A Modern Approach. 5th ed. South-Western, Mason, OH.

ГЛАВА 8. КОНТРФАКТИВНЫЕ СУЖДЕНИЯ: ГЛУБИННЫЙ АНАЛИЗ МИРОВ, КОТОРЫЕ МОГЛИ БЫ СУЩЕСТВОВАТЬ

Annotated Bibliography

The definition of counterfactuals as derivatives of structural equations was introduced by Balke and Pearl (1994a, 1994b) and was used to estimate probabilities of causation in legal settings. The relationships between this framework and those developed by Rubin and Lewis are discussed at length in Pearl (2000, Chapter 7), where they are shown to be logically equivalent; a problem solved in one framework would yield the same solution in another.

Recent books in social science (e.g., Morgan and Winship, 2015) and in health science (e.g., VanderWeele, 2015) are taking the hybrid, graph-counterfactual approach pursued in our book.

The section on linear counterfactuals is based on Pearl (2009, pp. 389–391), which also provides the solution to the problem posed in note 12. Our discussion of ETT is based on Shpitser and Pearl (2009). Legal questions of attribution, as well as probabilities of causation, are discussed at length in Greenland (1999), who pioneered the counterfactual approach to such questions. Our treatment of PN, PS, and PNS is based on Tian and Pearl (2000) and Pearl (2009, Chapter 9). A gentle approach to counterfactual attribution, including a tool kit for estimation, is given in Pearl, Glymour, and Jewell (2016). An advanced formal treatment of actual causation can be found in Halpern (2016).

Matching techniques for estimating causal effects are used routinely by potential outcome researchers (Sekhon, 2007), though they usually ignore the pitfalls shown in our education-experience-salary example. My realization that missing-data problems should be viewed in the context of causal modeling was formed through the analysis of Mohan and Pearl (2014).

Cowles (2016) and Reid (1998) tell the story of Neyman's tumultuous years in Lon*do*n, including the anec*do*te about Fisher and the wooden models. Greiner (2008) is a long and substantive introduction to "but-for" causation in the law. Allen (2003), Stott et al. (2013), Trenberth (2012), and Hannart et al. (2016) address the problem of attribution of weather events to climate change, and Hannart in particular invokes the ideas of necessary and sufficient probability, which bring more clarity to the subject.

References

Allen, M. (2003). Liability for climate change. Nature 421: 891–892. Balke, A., and Pearl, J. (1994a). Counterfactual probabilities: Computational methods, bounds, and applications. In Uncertainty in Artificial Intelligence 10 (R. L. de Mantaras and D. Poole, eds.). Morgan Kaufmann, San Mateo, CA, 46–54.

Balke, A., and Pearl, J. (1994b). Probabilistic evaluation of counterfactual queries. In Proceedings of the Twelfth National Conference on Artificial Intelligence, vol. 1. MIT Press, Menlo Park, CA, 230–237.

Cowles, M. (2016). Statistics in Psychology: An Historical Perspective. 2nd ed. Routledge, New York, NY.

Duncan, O. (1975). Introduction to Structural Equation Models. Academic Press, New York, NY.

Freedman, D. (1987). As others see us: A case study in path analysis (with discussion). Journal of Educational Statistics 12: 101–223. Greenland, S. (1999). Relation of probability of causation, relative risk, and *do*ubling *dos*e: A metho*do*logic error that has become a social problem. American Journal of Public Health 89: 1166–1169. Greiner, D. J. (2008). Causal inference in civil rights litigation. Harvard Law Review 81: 533–598.

Haavelmo, T. (1943). The statistical implications of a system of simultaneous equations. Econometrica 11: 1–12. Reprinted in D. F. Hendry and M. S. Morgan (Eds.), The Foundations of Econometric Analysis, Cambridge University Press, Cambridge, UK, 477–490, 1995.

Halpern, J. (2016). Actual Causality. MIT Press, Cambridge, MA. Hannart, A., Pearl, J., Otto, F., Naveu, P., and Ghil, M. (2016).

Causal counterfactual theory for the attribution of weather and climate-related events. Bulletin of the American Meteorological Society (BAMS) 97: 99–110.

Holland, P. (1986). Statistics and causal inference. Journal of the American Statistical Association 81: 945–960.

Hume, D. (1739). A Treatise of Human Nature. Oxford University Press, Oxford, UK. Reprinted 1888.

Hume, D. (1748). An Enquiry Concerning Human Understanding. Reprinted Open Court Press, LaSalle, IL, 1958.

Joffe, M. M., Yang, W. P., and Feldman, H. I. (2010). Selective ignorability assumptions in causal inference. International Journal of Biostatistics 6. *do*i:10.2202/1557-4679.1199.

Lewis, D. (1973a). Causation. Journal of Philosophy 70: 556–567. Re- printed with postscript in D. Lewis, Philosophical Papers, vol. 2, Oxford University Press, New York, NY, 1986.

Lewis, D. (1973b). Counterfactuals. Harvard University Press, Cambridge, MA.

Lewis, M. (2016). The Un*do*ing Project: A Friendship That Changed Our Minds. W. W. Norton and Company, New York, NY. Mohan, K., and Pearl, J. (2014). Graphical models for recovering probabilistic and causal queries from missing data. Proceedings of Neural Information Processing 27: 1520–1528.

Morgan, S., and Winship, C. (2015). Counterfactuals and Causal Inference: Methods and Principles for Social Research (Analytical Methods for Social Research). 2nd ed. Cambridge University Press, New York, NY.

Neyman, J. (1923). On the application of probability theory to agricultural experiments. Essay on principles. Section 9. Statistical Science 5: 465–480.

Pearl, J. (2000). Causality: Models, Reasoning, and Inference. Cambridge University Press, New York, NY.

Pearl, J. (2009). Causality: Models, Reasoning, and Inference. 2nd ed. Cambridge University Press, New York, NY.

Pearl, J., Glymour, M., and Jewell, N. (2016). Causal Inference in Statistics: A Primer. Wiley, New York, NY.

Reid, C. (1998). Neyman. Springer-Verlag, New York, NY.

Rubin, D. (1974). Estimating causal effects of treatments in ran*d*-

*o*mized and nonran*do*mized studies. Journal of Educational Psychol- ogy 66: 688–701.

Sekhon, J. (2007). The Neyman-Rubin model of causal inference and estimation via matching methods. In The Oxford Handbook of Political Metho*do*logy (J. M. Box-Steffensmeier, H. E. Brady, and D. Collier, eds.). Oxford University Press, Oxford, UK.

Shpitser, I., and Pearl, J. (2009). Effects of treatment on the treated: Identification and generalization. In Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence. AUAI Press, Montreal, Quebec, 514–521.

Stott, P. A., Allen, M., Christidis, N., *Do*le, R. M., Hoerling, M., Huntingford, C., Pardeep Pall, J. P., and Stone, D. (2013). Attri- bution of weather and climate-related events. In Climate Science for Serving Society: Research, Modeling, and Prediction Priorities (G. R. Asrar and J. W. Hurrell, eds.). Springer, *Do*rdrecht, Netherlands, 449–484.

Tian, J., and Pearl, J. (2000). Probabilities of causation: Bounds and identification. Annals of Mathematics and Artificial Intelligence 28: 287–313.

Trenberth, K. (2012). Framing the way to relate climate extremes to climate change. Climatic Change 115: 283–290.

VanderWeele, T. (2015). Explanation in Causal Inference: Methods for Mediation and Interaction. Oxford University Press, New York, NY.

ГЛАВА 9. ОПОСРЕДОВАНИЕ: В ПОИСКАХ МЕХАНИЗМА ДЕЙСТВИЯ

Annotated Bibliography

There are several books dedicated to the topic of mediation. The most up-to-date reference is VanderWeele (2015); MacKinnon (2008) also contains many examples. The dramatic transition from the statistical approach of Baron and Kenny (1986) to the counterfactual-based approach of causal mediation is described in Pearl (2014) and Kline (2015). Mc*Do*nald's quote (to discuss mediation, "start from scratch") is taken from Mc*Do*nald (2001).

Natural direct and indirect effects were conceptualized in Robins and Greenland (1992) and deemed problematic. They were later

formalized and legitimized in Pearl (2001), leading to the Mediation Formula.

In addition to the comprehensive text of VanderWeele (2015), new results and applications of mediation analysis can be found in De Stavola et al. (2015); Imai, Keele, and Yamamoto (2010); and Muthén and Asparouhov (2015). Shpitser (2013) provides a general criterion for estimating arbitrary path-specific effects in graphs.

The Mediation Fallacy and the fallacy of "conditioning" on a mediator are demonstrated in Pearl (1998) and Cole and Hernán (2002). Fisher's falling for this fallacy is told in Rubin (2005), whereas Rubin's dismissal of mediation analysis as "deceptive" is expressed in Rubin (2004).

The startling story of how the cure for scurvy was "lost" is told in Lewis (1972) and Ceglowski (2010). Barbara Burks's story is told in King, Montañez Ramírez, and Wertheimer (1996); the quotes from Terman and Burks's mother are drawn from the letters (L. Terman to R. Tolman, 1943).

The source paper for the Berkeley admissions para*do*x is Bickel, Hammel, and O'Connell (1975), and the ensuing correspondence between him and Kruskal is found in Fairley and Mosteller (1977).

VanderWeele (2014) is the source for the "smoking gene" example, and Bierut and Cesarini (2015) tells the story of how the gene was discovered.

The surprising history of tourniquets, before and during the Gulf War, is told in Welling et al. (2012) and Kragh et al. (2013). The latter article is written in a personal and entertaining style that is quite unusual for a scholarly publication. Kragh et al. (2015) describes the research that unfortunately failed to prove that tourniquets improve the chances for survival.

References

Baron, R., and Kenny, D. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. Journal of Personality and Social Psychology 51: 1173–1182.

Bickel, P. J., Hammel, E. A., and O'Connell, J. W. (1975). Sex bias in graduate admissions: Data from Berkeley. Science 187: 398–404.

Bierut, L., and Cesarini, D. (2015). How genetic and other biological factors interact with smoking decisions. Big Data 3: 198–202.

Burks, B. S. (1926). On the inadequacy of the partial and multiple correlation technique (parts I–II). Journal of Experimental Psychology 17: 532–540, 625–630.

Burks, F., to Mrs. Terman. (June 16, 1943). Correspondence. Lewis M. Terman Archives, Stanford University.

Ceglowski, M. (2010). Scott and scurvy. Idle Words (blog). Available at: http://www.idlewords.com/2010/03/scott_and_scurvy.htm (posted: March 6, 2010).

Cole, S., and Hernán, M. (2002). Fallibility in estimating direct effects. International Journal of Epidemiology 31: 163–165.

De Stavola, B. L., Daniel, R. M., Ploubidis, G. B., and Micali, N. (2015). Mediation analysis with intermediate confounding. American Journal of Epidemiology 181: 64–80.

Fairley, W. B., and Mosteller, F. (1977). Statistics and Public Policy. Addison-Wesley, Reading, MA.

Imai, K., Keele, L., and Yamamoto, T. (2010). Identification, inference, and sensitivity analysis for causal mediation effects. Statistical Science 25: 51–71.

King, D. B., Montañez Ramírez, L., and Wertheimer, M. (1996). Barbara Stoddard Burks: Pioneer behavioral geneticist and humanitarian. In Portraits of Pioneers in Psychology (C. W. G. A. Kimble and M. Wertheimer, eds.), vol. 2. Erlbaum Associates, Hillsdale, NJ, 212–225.

Kline, R. B. (2015). The mediation myth. Chance 14: 202–213.

Kragh, J. F., Jr., Nam, J. J., Berry, K. A., Mase, V. J., Jr., Aden, J. K., III, Walters, T. J., Dubick, M. A., Baer, D. G., Wade, C. E., and Blackbourne, L. H. (2015). Transfusion for shock in U.S. military war casualties with and without tourniquet use. Annals of Emergency Medicine 65: 290–296.

Kragh, J. F., Jr., Walters, T. J., Westmoreland, T., Miller, R. M., Mabry, R. L., Kotwal, R. S., Ritter, B. A., Hodge, D. C., Grey- danus, D. J., Cain, J. S., Parsons, D. S., Edgar, E. P., Harcke, T., Baer, D. G., Dubick, M. A., Blackbourne, L. H., Montgomery, H. R., Holcomb, J. B., and Butler, F. K. (2013). Tragedy into drama: An American history of tourniquet use in the current war. Journal of Special Operations Medicine 13: 5–25.

Lewis, H. (1972). Medical aspects of polar exploration: Sixtieth anniversary of Scott's last expedition. Journal of the Royal Society of Medicine 65: 39–42.

MacKinnon, D. (2008). Introduction to Statistical Mediation Analysis. Lawrence Erlbaum Associates, New York, NY.

Mc*Do*nald, R. (2001). Structural equations modeling. Journal of Consumer Psychology 10: 92–93.

Muthén, B., and Asparouhov, T. (2015). Causal effects in mediation modeling. Structural Equation Modeling 22: 12–23.

Pearl, J. (1998). Graphs, causality, and structural equation models. Sociological Methods and Research 27: 226–284.

Pearl, J. (2001). Direct and indirect effects. In Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence. Morgan Kaufmann, San Francisco, CA, 411–420.

Pearl, J. (2014). Interpretation and identification of causal mediation. Psychological Methods 19: 459–481.

Robins, J., and Greenland, S. (1992). Identifiability and exchangeability for direct and indirect effects. Epidemiology 3: 143–155.

Rubin, D. (2004). Direct and indirect causal effects via potential outcomes. Scandinavian Journal of Statistics 31: 161–170.

Rubin, D. (2005). Causal inference using potential outcomes: Design, modeling, decisions. Journal of the American Statistical Association 100: 322–331.

Shpitser, I. (2013). Counterfactual graphical models for longitudinal mediation analysis with unobserved confounding. Cognitive Science 37: 1011–1035.

Terman, L., to Tolman, R. (August 6, 1943). Correspondence. Lewis M. Terman Archives, Stanford University.

VanderWeele, T. (2014). A unification of mediation and interaction: A four-way decomposition. Epidemiology 25: 749–761.

VanderWeele, T. (2015). Explanation in Causal Inference: Methods for Mediation and Interaction. Oxford University Press, New York, NY.

Welling, D., MacKay, P., Rasmussen, T., and Rich, N. (2012). A brief history of the tourniquet. Journal of Vascular Surgery 55: 286–290.

ГЛАВА 10. БОЛЬШИЕ ДАННЫЕ, ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ И ВАЖНЫЕ ВОПРОСЫ

Annotated Bibliography

An accessible source for the perpetual free will debate is Harris (2012). The compatibilist school of philosophers is represented in the writings of Mumford and Anjum (2014) and Dennett (2003).

Artificial intelligence conceptualizations of agency can be found in Russell and Norvig (2003) and Wooldridge (2009). Philosophical views on agency are compiled in Bratman (2007). An intent-based learning system is described in Forney et al. (2017).

The twenty-three principles for "beneficial AI" agreed to at the 2017 Asilomar meeting can be found at Future of Life Institute (2017).

References

Bratman, M. E. (2007). Structures of Agency: Essays. Oxford University Press, New York, NY.

Brockman, J. (2015). What to Think About Machines That Think. HarperCollins, New York, NY.

Dennett, D. C. (2003). Free*do*m Evolves. Viking Books, New York, NY.

Forney, A., Pearl, J., and Bareinboim, E. (2017). Counterfactual data-fusion for online reinforcement learners. Proceedings of the 34th International Conference on Machine Learning. Proceedings of Machine Learning Research 70: 1156–1164.

Future of Life Institute. (2017). Asilomar AI principles. Available at: https://futureoflife.org/ai-principles (accessed December 2, 2017). Harris, S. (2012). Free Will. Free Press, New York, NY.

Mumford, S., and Anjum, R. L. (2014). Causation: A Very Short Introduction (Very Short Introductions). Oxford University Press, New York, NY.

Russell, S. J., and Norvig, P. (2003). Artificial Intelligence: A Modern Approach. 2nd ed. Prentice Hall, Upper Saddle River, NJ. Wooldridge, J. (2009). Introduction to Multi-agent Systems. 2nd ed. John Wiley and Sons, New York, NY.

Издание для досуга
демалысқа арналған баспа

12+

Перл Джудиа, Маккензи Дана

ДУМАЙ «ПОЧЕМУ?»
Причина и следствие как ключ к мышлению

Переводчики *Т. Мамедова и М. Антипин*

Ответственный редактор *А. Ходякова*
Менеджер проекта *В. Живина*
Технический редактор *Н. Чернышева*
Корректор *М. Маркова*
Дизайн *А. Закопайко и О. Жукова*
Компьютерная верстка *Н. Шаповалова*

Подписано в печать 26.09.2022. Формат 60x90/16.
Усл. печ. л. 28.0. Печать офсетная. Гарнитура CharterITC.
Бумага офсетная. Тираж 2500 экз. (Trend book) Заказ №
Тираж 2000 экз. (Власть и успех) Заказ №

Произведено в Российской Федерации. Изготовлено в 2022 г.
Оригинал-макет подготовлен редакцией «Времена»,
импринт «Альфа»

Изготовитель: ООО «Издательство АСТ»
129085, Российская Федерация, г. Москва, Звездный бульвар,
д. 21, стр. 1, комн. 705, пом. I, этаж 7
Наш сайт: WWW.AST.RU E-mail: ask@ast.ru
Общероссийский классификатор продукции ОК-034-2014
(КПЕС 2008); 58.11.1 - книги, брошюры печатные

«Баспа Аста» деген ООО
129085, г. Мәскеу, Жулдызды гүлзар, д. 21, 1 құрылым, 705 бөлме, пом. 1,
7-қабат. Біздің электрондық мекенжайымыз : www.ast.ru
E-mail: ask@ast.ru Интернет-магазин: www.book24.kz
Интернет-дүкен: www.book24.kz
Импортер в Республику Казахстан и Представитель по приему претензий
в Республике Казахстан — ТОО РДЦ Алматы, г. Алматы.
Қазақстан Республикасына импорттаушы және Қазақстан Республикасында
наразылықтарды қабылдау бойынша өкіл -«РДЦ-Алматы» ЖШС, Алматы қ.,
Домбровский көш., 3«а», Б литері офис 1. Тел.: 8 (727) 251 59 90,91 ,
факс: 8 (727) 251 59 92 ішкі 107; E-mail: RDC-Almaty@eksmo.kz ,
www.book24.kz Тауар белгісі: «АСТ» Өндірілген жылы: 2022
Өнімнің жарамдылық; мерзімі шектелмеген.
Сертификация қарастырылмаған