

Брайан
Кристиан

БЕССОВЕСТНАЯ МАШИНА

МОЖНО ЛИ НАУЧИТЬ ИИ ЭМПАТИИ,
СОСТРАДАНИЮ И ДРУГИМ
ЧЕЛОВЕЧЕСКИМ ЦЕННОСТЯМ?



Исследование ИИ
от автора бестселлера
«Алгоритмы для жизни»



БОМБОРА
ИЗДАТЕЛЬСТВО

Москва

УДК 004.8
ББК 32.813
К82

Brian Christian
The Alignment Problem: Machine Learning and Human Values
Copyright © 2020 by Brian Christian. All rights reserved.

В книге приводятся примеры, в том числе связанные с проявлениями нетолерантности по расовому, гендерному и другим признакам. Все подобные случаи и результаты исследований представлены исключительно в научных и образовательных целях, для анализа существующих проблем в области искусственного интеллекта, а также для иллюстрации этических вызовов, с которыми сталкиваются современные технологии.

Автор и издательство не поддерживают и не оправдывают любые формы дискриминации, предвзятости или неравенства. Описанные в книге примеры не являются призывом к действию или выражением личной позиции, а служат материалом для критического осмысления и обсуждения путей совершенствования алгоритмов и систем ИИ.

Чтение книги предполагает осознанное отношение к этим вопросам и стремление к развитию более справедливых и этичных технологий будущего.

Кристиан, Брайан.

К82 Бессовестная машина. Можно ли научить ИИ эмпатии, состраданию и другим человеческим ценностям? / Кристиан Брайан ; [перевод с английского В. В. Краснянской]. — Москва : Эксмо, 2026. — 512 с. — (Будущее: прогнозы и технологии).

ISBN 978-5-04-238776-0

«Бессовестная машина» Кристиана Брайана — это глубокое исследование одной из самых острых проблем современности: как научить искусственный интеллект действовать в соответствии с человеческими ценностями и этикой. Автор рассматривает случаи, когда ИИ усиливает социальные предубеждения, принимает несправедливые решения и выходит из-под контроля. Книга сочетает в себе исторический анализ, находки передовой науки и эксперименты. Кристиан показывает, как развитие ИИ переплетается с психологией, философией и культурой и почему решение проблемы выравнивания станет определяющим для будущего человечества. Это не только предупреждение о рисках, но и вдохновляющий рассказ о попытках сделать технологии безопасными и полезными для всех.

УДК 004.8
ББК 32.813

ISBN 978-5-04-238776-0

© Краснянская В.В., перевод на русский язык, 2026
© Оформление. ООО «Издательство «Эксмо», 2026

ОГЛАВЛЕНИЕ

Пролог	10
Введение	14

Часть 1. Предсказание

1. Представление	30
2. Справедливость	74
3. Прозрачность	115

Часть 2. Агенты

4. Подкрепление	162
5. Формирование навыков	201
6. Любопытство	235

Часть 3. Нормативность

7. Имитация	270
8. Логический вывод	310
9. Неопределенность	342
Заключение	380
Благодарности	403
Алфавитный указатель	408
Примечания	415

*Посвящается Питеру,
который меня убедил.*

*И всем, кто выполняет
свою работу*

Помню, в 2000 году я услышал, как Джеймс Мартин, руководитель программы Viking, отправивший беспилотные космические аппараты на Марс, говорит о своей работе. Ему как инженеру-разработчику космических кораблей предстояло совершить посадку не на Марс, а на модель красной планеты, придуманную геологами.

— ПИТЕР НОРВИГ¹

Мир — это лучшая модель его самого.

— РОДНИ БРУКС²

Все модели неверны.

— ДЖОРДЖ БОКС³

ПРОЛОГ

1 935 год, Детройт. Уолтер Питтс во весь опор мчался вниз по улице, спасаясь от преследователей.

Он добрался до публичной библиотеки и спрятался там. Убежище оказалось настолько хорошим, что Уолтера не нашли даже работники библиотеки и закрыли помещение на ночь. Питтс оказался запертым внутри⁴.

Одна из книг на полках показалась ему интересной, и он начал читать. За три дня Уолтер прочел ее от корки до корки.

В книге было две тысячи страниц, посвященных формальной логике; что примечательно, доказательство того, что $1+1=2$, появилось только на 379 странице⁵. Питтс решил написать одному из авторов — британскому философу Бертрану Расселу, — поскольку ему показалось, что в книге допущено немного ошибок.

Через несколько недель Уолтер получил письмо из Англии. Ему ответил Бертран Рассел. Ученый поблагодарил своего читателя и пригласил поступить в аспирантуру в Кембридже⁶.

К несчастью, Уолтеру Питтсу пришлось отклонить это предложение — молодому человеку было всего двенадцать лет и он учился в седьмом классе.

Три года спустя Питтс узнал, что Рассел приезжает читать лекции в Чикаго, и сбежал из дома, чтобы побывать на них. Обратно Уолтер уже не вернулся.

* * *

На лекции Рассела Питтс познакомился с еще одним подростком, которого звали Джерри Летвин. Питтса интересовала только логика. Летвина — поэзия и — несколько сильнее — медицина⁷. Молодые люди стали неразлучными друзьями.

Питтс слонялся по кампусу университета Чикаго, бессистемно посещая занятия; аттестата об окончании средней школы у него так и не было, и официально в списках студентов он не числился. Один из курсов вел прославленный немецкий логик Рудольф Карнап. Как-то раз Питтс зашел в его кабинет и заявил, что нашел несколько «недочетов» в последней книге Карнапа. Тот отнесся к этим словам скептически, но все-таки сверился с текстом. Питтс, конечно же, оказался прав. Некоторое время они беседовали, потом Питтс вышел из кабинета, так и не назвав своего имени. Несколько месяцев Карнап расспрашивал всех вокруг о «разносчике газет, который разбирается в логике»⁸. В конце концов Карнап нашел Питтса и, как потом часто случалось в академической карьере молодого человека, стал его покровителем, убедив администрацию университета дать юноше хоть какую-то неквалифицированную работу, чтобы обеспечить его постоянным доходом.

Шел 1941 год. Летвин, который по-прежнему считал себя в первую очередь поэтом, все же поступил в медицинскую школу университета Иллинойса и обнаружил, что ему предстоит работать под руководством гениального нейрофизиолога Уоррена Маккаллока, недавно прибывшего из Йеля. Однажды Летвин предложил Питтсу встретиться с Маккал-

локом. Летвину уже был двадцать один год, но он все еще жил с родителями. Питтсу было семнадцать, и дома у него не было⁹. Маккаллоу и его жена приютили обоих молодых людей.

В течение следующего года Маккаллоу приходил домой по вечерам, и они с Питтсом, который был немногим старше детей ученого, засиживались до полуночи за разговорами. Из них получилась отличная интеллектуальная команда: пользующийся уважением нейрофизиолог на взлете своей карьеры и юное дарование в области логики. Один жил практикой — в мире нервной системы и невротических расстройств, другому была известна только теория — мир символов и доказательств. Оба стремились исключительно к познанию сущности истины: что это такое и как мы ее узнаем. Точкой приложения сил, тем, что находится точно на перекрестке двух несопоставимых миров, оказался мозг.

В начале 1940-х годов уже было известно, что мозг состоит из связанных между собой нейронов, а также о том, что у каждого нейрона есть «входные» каналы (дендриты) и «выходные» (аксоны). Когда входящие в нейрон импульсы достигают определенного порога, нейрон, в свою очередь, начинает передавать сигнал. Маккаллоу и Питтс немедленно восприняли это как логическое положение: импульс или его отсутствие можно обозначить как «включено — выключено», «да — нет», «правда — ложь»¹⁰.

Они установили, что нейрон с достаточно низким порогом восприимчивости, который реагирует на *любые* входящие сигналы, функционирует как физическое воплощение логической операции *или*. Нейрон с достаточно высоким порогом восприимчивости, который «загорается», только если задействованы *все* его входящие каналы, — это физическое воплощение логической операции *и*. Тогда исследователи начали понимать, что любое логическое умозаключение подобная «нейронная сеть» может реализовать, если она, конечно, правильно спроектирована.

В течение нескольких месяцев нейрофизиолог средних лет и логик-подросток написали статью. Они назвали ее «Логическое исчисление идей, относящихся к нервной активности».

«Поскольку нервная активность подчиняется закону „все или ничего“, — писали они, — то нейронные события и соотношения между ними можно изучать средствами логики высказываний. Оказывается, что поведение любой сети может быть описано в этих терминах... Для всякого логического выражения, удовлетворяющего некоторым условиям, можно найти сеть, имеющую описываемое этим выражением поведение»¹.

Статья была опубликована в 1943 году в журнале *Bulletin of Mathematical Biophysics*. К досаде Летвина, она слабо повлияла на биологическое научное сообщество¹¹. К разочарованию Питтса, нейрофизиологические работы 1950-х годов, а в особенности примечательное исследование зрительных нервов лягушки, проведенное никем иным как его лучшим другом Джерри Летвином, показали, что нейроны имеют гораздо более сложные связи, чем простые цепочки истинных и ложных положений, какими себе их представлял Уолтер. Возможно, алгебра высказываний и ее конъюнкции, дизъюнкции и отрицания, в конце концов, не имели отношения к языку мозга или, по крайней мере, представляли его в более опосредованной форме. Эти неясности огорчали Питтса.

Но влияние статьи, ставшей плодом долгих ночных разговоров в доме Маккаллока, оказалось огромным, хотя и не совсем таким, как представляли ее авторы. Она легла в основу совершенно новой отрасли, ставящей цель создать в реальности механизмы, являющиеся той самой упрощенной копией нейронов, и посмотреть, что такие «механические мозги» смогут делать¹².

¹ Мак-Каллок У. С., Питтс В. Логическое исчисление идей, относящихся к нервной активности // Нейрокомпьютер. — 1992. — № 3, 4. — с. 40–53.

ВВЕДЕНИЕ

Летом 2013 года в открытом блоге компании Google появился совершенно безобидный пост под названием «Изучение смысла, скрывающегося за словами»¹³.

Начинался он так: «Сегодня компьютеры не слишком хорошо понимают человеческий язык. И хотя современное состояние технологий в этом плане все еще оставляет желать лучшего, мы добились значительного прогресса, используя последние достижения машинного обучения и методы обработки естественного языка».

Сотрудники Google «скормили» искусственной нейросети, созданной по образу биологических систем, огромные базы данных естественного языка. Эти тексты, собранные с из новостных статей и Интернета, по объему в *тысячи* раз превосходили все, что использовалось ранее. Системе позволили самостоятельно анализировать предложения в поисках взаимосвязей и зависимостей между словами.

С помощью метода так называемого обучения без учителя система начала выявлять закономерности. К примеру, она определила, что слово «Пекин» (что бы оно ни значило для программы) находится в таких же отношениях со словом

«Китай» (что бы за ним ни стояло), как слово «Москва» — со словом «Россия».

Можно ли назвать это «пониманием» — вопрос для философов, однако трудно поспорить с тем, что алгоритм уловил нечто очень важное в самом процессе чтения и осмысления текста.

Поскольку программа преобразовывала прочитанные слова в числовые значения, называемые векторами, Google дала ей имя Word2Vec и опубликовала как программное обеспечение с открытым исходным кодом.

С точки зрения математики векторы обладают набором полезных свойств, позволяющих оперировать ими как обычными числами: их можно складывать, вычитать и умножать. Вскоре исследователи обнаружили нечто удивительное и неожиданное. Они назвали это явление «лингвистическими закономерностями в непрерывном векторном пространстве представления слов»¹⁴, но объяснить его можно гораздо проще. Поскольку Word2Vec превращает слова в векторы, к ним можно применять *математические операции*.

Например, если вы напишете «Китай+река», то получите «Янцзы». Если введете «Париж-Франция+Италия», результатом будет «Рим». А из выражения «король-мужчина+женщина» получится «королева».

Результаты оказались поразительными. Технологию Word2Vec начали применять в «Google Переводчике» и поисковых алгоритмах компании, что подтолкнуло к ее внедрению во множество других сервисов, включая платформы для найма и поиска работы. Эта система стала одним из главных инструментов для нового поколения лингвистов, изучающих большие данные в университетах по всему миру.

В течение двух лет никто не замечал в этом проблемы.

В ноябре 2015 года аспирант Бостонского университета Толга Болукбаси вместе со своим научным руководителем пришел на неформальную пятничную встречу в Microsoft Research. Потягивая вино и общаясь, они вместе с исследователем из Microsoft Адамом Калаи открыли ноутбуки и начали экспериментировать с Word2vec.

«Мы играли с векторами слов и ради интереса вводили случайные комбинации, — рассказывал Болукбаси. — Я сидел за своим компьютером, Адам присоединился ко мне»¹⁵. А затем кое-что произошло.

Они набрали:

«врач – мужчина + женщина»

И в ответ получили:

«медсестра»

«В этот момент мы были просто в шоке и поняли, что есть проблема, — рассказывал Калаи. — Мы копнули глубже и выяснили, что все обстоит гораздо хуже»¹⁶.

Исследователи попробовали еще одно сочетание:

«лавочник – мужчина + женщина»

Ответом стало:

«домохозяйка»

Еще одно выражение:

«программист – мужчина + женщина»

дало тот же ответ:

«домохозяйка»

К этому моменту все разговоры в комнате стихли, а вокруг монитора собралась целая группа людей. Как говорит Болукбаси: «Мы все одновременно осознали, что что-то пошло не так».

* * *

В органах правосудия по всей стране все больше и больше судей полагается на алгоритмические инструменты анализа рисков, принимая решения о размере залога или самом факте содержания обвиняемого под стражей до рассмотрения дела. Комиссии по условно-досрочному освобождению используют их, освобождая заключенных до истечения срока наказания или отказывая в этой привилегии. Один из самых популярных таких инструментов был разработан мичиганской фирмой Northpointe и известен как программа прогноза криминального рецидива Correctional Offender Management Profiling for Alternative Sanctions, сокращенно COMPAS¹⁷. Она использовалась в Калифорнии, Флориде, Нью-Йорке, Мичигане, Висконсине, Нью-Мексико и Вайоминге, алгоритмически оценивая степени рисков — рецидива в целом, повторного совершения тяжкого преступления или неявки в суд — по шкале от 1 до 10.

Как ни удивительно, подобные средства зачастую применялись на территории всего штата без официальных проверок¹⁸. COMPAS — патентованный инструмент с закрытым исходным кодом, так что ни прокуроры, ни адвокаты, ни судьи точно не знали, как работает его модель.

В 2016 году группа журналистов-аналитиков данных из некоммерческой организации ProPublica под руководством Джулии Энгвин решила тщательнее изучить COMPAS. Запросив судебные протоколы округа Броуард, штат Флорида, журналисты получили материалы дел и оценки рисков