Брайан Кристиан

БЕССОВЕСТНАЯ МАШИНА

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

18+

Исследование ИИ
от автора бестселлера
«Алгоритмы для жизни»

БОМБОРА
ИЗДАТЕЛЬСТВО

Брайан Кристиан

Бессовестная машина. Можно ли научить ИИ эмпатии, состраданию и другим человеческим ценностям?

Серия «Будущее: прогнозы и технологии»

http://www.litres.ru/pages/biblio_book/?art=74136731
Бессовестная машина. Можно ли научить ИИ эмпатии, состраданию и другим человеческим ценностям?:
ISBN 978-5-04-250566-9

Аннотация

Хорошо это или плохо, но история человечества в этом столетии плотно связана с созданием и поддержанием систем искусственного интеллекта. Как ученики волшебника, мы обнаружили, что являемся всего лишь одной движущей силой среди многих в мире, заполненном волшебными метлами. Как именно мы собираемся обучать машины? И чему?

Брайан Кристиан исследует ключевой вызов эпохи ИИ: как научить машины понимать и разделять человеческие ценности. От алгоритмов, усиливающих предвзятость, до непредсказуемых

последствий «умных» систем, книга показывает, что просто создать ИИ – недостаточно. Необходимо обеспечить его «согласование» с нашими намерениями. Кристиан предлагает увлекательный взгляд на историю машинного обучения и призывает к более гуманному подходу к разработке ИИ, где человеческое мнение играет центральную роль.

Точки пересечения ИИ и человека:

1. Справедливость и равенство: невозможно одновременно учитывать разные критерии справедливости из-за существующих социальных различий.

2. Предупреждение рисков: новые технологии обучения и адаптации моделей должны сглаживать негативные последствия ИИ.

3. Прозрачность и объяснимость: необходимо понимать принципы работы ИИ, чтобы не попадать в опасные ситуации.

4. Коллективная ответственность: нужен междисциплинарный подход к созданию и регулированию ИИ, учитывающий мнения представителей разных областей знания.

В формате PDF A4 сохранен издательский макет книги.

Содержание

Кристиан Брайан
Бессовестная машина
Можно ли научить ИИ эмпатии, состраданию и другим человеческим ценностям?

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 Норвиг [1]

Мир – это лучшая модель его самого.
– Родни Брукс [2]

Все модели неверны.
– Джордж Бокс [3]

Пролог

1935 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 внутри [4].

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 странице [5]. Питтс решил

---

[4] Информация о жизни Уолтера Питтса чрезвычайно скудна. Я нашел очень мало первоисточников, в основном, это были письма Питтса к Уоррену Маккаллоку, которые доступны в архиве Маккаллока в Американском философском обществе в Филадельфии. Я благодарен его персоналу за помощь. Другие материалы были взяты из устных рассказов современников Питтса, а именно Джерома (Джерри) Леттвина в книге Anderson and Rosenfeld, *Talking Nets,* а также в эссе и воспоминаниях в книге McCulloch, *The Collected Works of Warren S. McCulloch.* Другие рассказы о жизни Питтса можно найти в книгах Smalheiser, Walter Pitts; Easterling, Walter Pitts и Gefter, *The Man Who Tried to Redeem the World with Logic*. Более подробные детали есть в биографиях Маккаллока, Норберта Винера и других кибернетиков, например Heims, John von Neumann and Norbert Wiener и The Cybernetics Group, and Conway and Siegelman, *Dark Hero of the Information Age*.

[5] *Уайтхед А., Рассел Б.* Основания математики: В 3 т. / Под ред. Г. П. Яровой, Ю. Н. Радаева. – Самара: Самарский университет, 2005–2006.

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 Кембридже [6].

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 – медицина [7]. 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 сту-

---

[6] Благодарю сотрудников архива Бертрана Рассела в Макмастерском университете за помощь в поисках копии этого письма; к сожалению, она не сохранилась.

[7] Anderson and Rosenfeld, *Talking Nets*.

дентов 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 Маккаллоком. Летвину уже был двадцать один год, но он все еще жил с родителями. Питтсу было семнадцать, и дома у него

---

[8] Anderson and Rosenfeld, *Talking Nets*. Видимо, имелась в виду книга Карнапа «Преодоление метафизики логическим анализом языка» (*Logische Syntax der Sprache*), хотя некоторые источники называют «Логическую структуру мира» (*Der logische Aufbau der Welt*).

не было [9]. Маккаллок и его жена приютили обоих молодых людей.

В течение следующего года Маккаллок 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. Маккаллок и Питтс немедленно восприняли это как логическое положение: импульс или его

---

[9] В зависимости от точной даты их знакомства, возможно, Питтсу уже было восемнадцать (а Летвину двадцать). Маккаллок пишет: «В 1941 году я представил свои размышления о потоке информации через ряды нейронов на семинаре Рашевского на факультете математической биологии Чикагского университета и познакомился с Уолтером Питтсом, которому тогда было около семнадцати». См. McCulloch, *The Collected Works of Warren S. McCulloch,* с. 35–36.

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 пове-

[10] Некоторые истоки этих мыслей предвосхищают совместную работу Маккаллока и Питтса; см. McCulloch, *Recollections of the Many Sources of Cybernetics*.

дение» [11].

Статья была опубликована в 1943 году в журнале *Bulletin of Mathematical Biophysics*. К досаде Летвина, она слабо повлияла на биологическое научное сообщество [12]. 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»

---

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

[12] См. Piccinini, *The First Computational Theory of Mind and Brain*, and Lettvin, Introduction to McCulloch, *The Collected Works of Warren S. McCulloch*.

смогут делать [13].

[13] В докладе Джона фон Неймана об ЭДВАК (электронном автоматическом вычислителе с дискретными переменными) – первом в истории описании компьютера с хранимой в памяти программой – на 101 страницу текста приходится лишь одна ссылка: на статью Маккаллока и Питтса 1943 года. (См. Neumann, *First Draft of a Report on the EDVAC*. К слову, в фамилии Маккаллока фон Нейман сделал опечатку.) Фон Нейман был увлечен их доказательствами и в разделе «Нейронные аналогии» рассматривал практическое применение вычислительных устройств, появление которых предвидел. «Очевидно, что эти нейронные функции в упрощенном виде могут имитироваться с помощью телеграфных реле или вакуумных электронных ламп, – писал он. – Поскольку соединения этих ламп должны передавать числа посредством цифр, вполне естественно использовать арифметическую систему, где цифры имеют два значения. А это предполагает использование бинарной системы». Все мы знаем, во что превратились эти бинарные машины с хранимыми в памяти программами, построенные на базе логических элементов. Это компьютеры, которые распространились настолько, что сегодня их количество превышает численность людей на планете.И, тем не менее, эта архитектура, вдохновленная строением головного мозга, быстро вышла за рамки «нейронной аналогии». Многие задавались вопросом: могут ли существовать машины, более близкие по своей архитектуре к мозгу? Такие, где использовался бы не один-единственный процессор, которому на огромной скорости поочередно «скармливают» четкие логические инструкции, а распределенная сеть из относительно простых, однотипно действующих ячеек, чья совокупная эффективность превышает простую сумму действий ее базовых компонентов. Возможно, даже не полностью бинарная, а наделенная той долей неупорядоченности, которую взял на вооружение Летвин, но которой избегал Питтс. Время от времени появлялись специализированные аппаратные средства для параллельной обработки нейронных сетей (включая перцептрон Mark I Фрэнка Розенблатта), но, как правило, это были уникальные, штучные устройства. Настоящая аппаратная революция, сделавшая возможным масштабное параллельное обучение нейросетей на базе графических процессоров (GPU), произошла лишь несколько десятилетий спустя – в середине 2000-х годов.

Введение

Летом 2013 года в открытом блоге компании Google появился совершенно безобидный пост под названием «Изучение смысла, скрывающегося за словами» [14].

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» –

---

[14] Mikolov, Sutskever, and Le, *Learning the Meaning Behind Words*.

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» [15], 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 по-

[15] Mikolov, Yih, and Zweig, *Linguistic Regularities in Continuous Space Word Representations*.

исковых 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 интересна вводили случайные комбинации, – рассказывал Болукбаси. – Я сидел за своим компьютером, Адам присоединился ко мне» [16]. 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» [17].

Исследователи попробовали еще одно сочетание:

---

[16] Интервью автора с Толгой Болукбаси, 11 ноября 2016.

[17] Интервью автора с Адамом Калаи, 4 апреля 2018.

«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

COMPAS [18]. 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 проверок [19]. 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 примерно семи тысяч обвиняемых, арестованных в период с 2013 по 2014 годы.

Поскольку расследование велось в 2016 году, члены команды ProPublica могли сыграть роль оракулов. Просматривая информацию за два предыдущих года, журналисты уже знали, как повели себя эти люди, нарушили они снова закон

---

[18] В январе 2017 года компания *Northpointe* слилась с *CourtView Justice Solutions* и *Constellation Justice Systems*. Они переименовали себя в *equivant* (со строчной буквы) и открыли штаб-квартиру в Огайо.

[19] «Эти проверки зачастую проводят те же самые люди, которые разработали инструмент» (Desmarais and Singh, *Risk Assessment Instruments Validated and Implemented in Correctional Settings in the United States*).

или нет. Было задано два простых вопроса: верно ли модель предсказывает, кто из обвиняемых наиболее опасен, и не отдает ли она предпочтение какой-то группе или не действует ли она против одной из категорий осужденных?

Заподозрить неладное можно было с первого взгляда на данные. Например, журналисты обнаружили, что двое обвиняемых получили одинаковые сроки за один и тот же серьезный тип преступления. Дилана Фьюгета, уже задерживали за попытку кражи. Бернарда Пакера обвиняли в ненасильственном сопротивлении аресту. Риск рецидива для белокожего Фьюгета программа оценила как 3 из 10. У чернокожего Пакера оценка рисков рецидива составила 10 из 10.

В 2016 году уже известно, что Фьюгет был трижды задержан за разные противоправные нарушения. В то же время Пакер не совершил ни одного преступления.

В другом случае журналисты сопоставили двух осужденных, получивших одинаковые сроки за мелкую кражу. Вернон Пратер привлекался трижды – за два вооруженных ограбления и одну попытку. Бриша Борден, будучи подростком, имела четыре приво́да. Риск рецидива у белого Пратера программа вновь оценила на 3 из 10, а у чернокожей Борден – на 8 из 10.

С позиции 2016 года команда Эгвин увидела, что Пратер в итоге получил еще один срок за кражу в особо крупных размерах и отправился в тюрьму на восемь лет. Борден же более ни за какие правонарушения не привлекалась.

Осужденные и сами недоумевают от подобных оценок. Джеймс Ривелли, белый мужчина, был арестован за мелкую магазинную кражу. Риск рецидива был оценен на 3 из 10, несмотря на то что ранее он уже обвинялся в физическом насилии при отягчающих обстоятельствах и имел несколько сроков за кражи. «Я провел пять лет в тюрьме штата Массачусетс, – рассказывал он репортеру. – Удивительно, что баллы такие низкие».

Статистический анализ подтвердил, что такая несоразмерность возникает систематически [20]. Вышла статья, основную мысль которой можно свести к двум предложениям: «Это программное обеспечение используется по всей стране, чтобы предсказывать преступления. И оно необъективно к чернокожим».

Но были и те, кто не был так уверен, и доклад ProPublica 2016 года вызвал целую бурю споров не только о COMPAS и алгоритмическом анализе рисков в целом, но и о самом понятии справедливости. Как именно мы должны определять в терминах статистического и вычислительного подходов принципы, права и идеалы, сформулированные законом?

Когда председатель Верховного суда США Джон Робертс годом позже посетил Политехнический институт Ренсселера, президент университета Ширли Энн Джексон спросила его:

---

[20] Angwin et al., *Machine Bias*.

– Как вы считаете, настанет день, когда умные машины с искусственным интеллектом будут участвовать в судебном следствии или, что еще более сомнительно, помогать выносить вердикт?

– Этот день уже настал, – ответил он [21].

* * *

Той же осенью Дарио Амодей приехал в Барселону на конференцию по нейросетевым системам обработки информации (Neural Information Processing Systems conference, сокращенно NeurIPS). Это самое крупное событие в отрасли ИИ, которое разрослось от нескольких сотен участников в 2000-е годы до тринадцати тысяч в наши дни. (Организаторы отмечают, что, если она будет продолжать расти такими же темпами, как в последние десять лет, к 2035 году на ней соберется *все население планеты*.) [22] Но именно в тот день ум Амодея был занят не «порядком сканирования в выборке Гиббса», не «регуляризацией потерь на наблюдениях с шумом Рademахера» и не «минимизацией со-

[21] Rensselaer Polytechnic Institute, *A Conversation with Chief Justice John G. Roberts, Jr.*

[22] Эту шутку придумал руководитель программы Сэми Бенджио во время вступительного слова на конференции 2017 года, см. https://media.nips.cc/Conferences/NIPS2017/Eventmedia/opening_remarks.pdf. Тринадцать тысяч участников посетили конференцию в 2019 году, см. https://huyenchip.com/2019/12/18/key-trends-neurips-2019.html

жалений в рефлексивных банаховых пространствах», и, если уж на то пошло, не представленным в соседней комнате тематическим докладом Толги Болукбаси о гендерном неравенстве в Word2Vec [23].

Он смотрел на лодку, которую охватил пожар.

Амодей наблюдал, как она заходит в маленький порт, делает несколько кругов и ее корма ударяется о каменный причал. Двигатель загорелся. Лодка продолжала бешено кружиться, брызги воды погасили пламя. Затем судно врезалось в бок буксирного катера, и пожар снова разгорелся. А потом лодка опять крутилась у причальной стены.

Это все происходило, потому что Амодей, по-видимому, приказал лодке вести себя именно так. Лодка делала то, что он велел. Но не то, что он *имел в виду*.

Амодей – один из исследователей, работавших в проекте Universe. Он член команды, разрабатывавшей один ИИ общего назначения, нужный для решения широкого круга задач и способный играть в сотни различных компьютерных игр не хуже человека. Эта трудная задача в сообществе, работающем с искусственным интеллектом, воспринималась как что-то вроде священного Грааля.

«Я просто запустил несколько этих сред, – рассказывал мне Амодей, – и подключался к ним через VPN [24], чтобы

---

[23] Bolukbasi et al., *Man Is to Computer Programmer as Woman Is to Homemaker?*

[24] **VPN** (VPN-сервис) – С ноября 2022 г. на территории Российской Федерации запрещено распространять информацию о VPN-сервисах с целью доступа к

посмотреть, как идут дела. Там были автомобильные гонки, с которыми все было в порядке, еще что-то вроде гонки на грузовиках и та самая гонка на *лодках*».

Амодей замолк на минуту, а потом продолжил: «Я смотрел на них и думал: „Эта лодка, что, плавает кругами? Что, черт побери, вообще происходит?“» [25] Лодка не просто двигалась случайным образом; она не «одичала» и не потеряла управление. На самом деле, все было наоборот. Она была *запрограммирована* на такое поведение. С точки зрения компьютера, он нашел идеальную стратегию и воплотил ее в жизнь до последней буквы. Однако смысла в ней не было никакого.

«В конце концов я посмотрел на вознаграждение», – рассказывал Амодей.

Он допустил старую и хорошо известную ошибку: «поощрение А в надежде на В» [26]. Чего хотел исследователь, так это научить машину выигрывать лодочную гонку. Но это трудно выразить четко. Амодею нужно было найти способ формализовать такие сложные понятия, как место на дорожке, круги, положение среди других лодок и так далее. Вместо этого

запрещенному контенту. Научная, научно-техническая и статистическая информация о VPN-сервисах для обхода блокировок признана запрещенной в России, исключением является информация о VPN для обеспечения защищенного удаленного доступа. – Прим. ред.

[25] Интервью автора с Дарио Амодеем 24 апреля 2018 года.

[26] Эта хорошо запоминающаяся формулировка взята из классической статьи Керра *On the Folly of Rewarding A, While Hoping for B.*

он использовал то, что показалось разумной заменой: очки. Машина нашла лазейку – маленькую гавань, где можно было пополнять запасы топлива и полностью игнорировать гонку, нарезать круги и *вечно* набирать очки.

«И, конечно же, в какой-то мере в этом был виноват я, – сказал Амодей. – Просто запустил все эти игры, не обратив должного внимания на целевые функции… В других играх нужно проходить гонку, чтобы набрать очки. Их начисляют за пользование заправками, которые расставлены вдоль дороги… В десяти других случаях очки зарабатывались победой. Но не в одиннадцатом [27].

Люди критиковали такой подход, говоря: „Ты получил именно то, о чем просил“. Это словно сказать: „Твое оптимальное решение не подразумевало финиша в гонке“. И я могу ответить на это только одно, – в этом месте он сделал паузу, – все так и есть».

Амодей разместил видеоролик с этим эпизодом на канале своей группы в корпоративном мессенджере Slack. Видео немедленно посчитали «уморительным» все, кто его видел. Эта мультяшная плохая комедия, безусловно, такой и была. Но для Амодея, который руководил командой по безопасности ИИ в исследовательской лаборатории OpenAI в Сан-Франциско, в этом эпизоде имелся куда более глубокий смысл. В каком-то смысле *именно такого поведения* он

[27] Официальное сообщение компании Open AI об инциденте с гоночной лодкой можно увидеть в статье Clark and Amodei, *Faulty Reward Functions in the Wild*.

всегда и боялся.

На самом деле смысл игры, в которую они с коллегами-исследователями играли, не в том, чтобы победить в гонках нарисованных лодочек; это попытка научить все более многофункциональные системы ИИ делать то, что мы хотим, – особенно когда трудно изложить четко и полно, что мы хотим и чего *не хотим*.

Сценарий с лодкой – это, конечно же, просто разминка, тренировка. Весь понесенный ущерб остался исключительно виртуальным. Но это тренировка в игре, которая на деле вовсе не игра. Начиналось с нескольких робких заявлений и постепенно захватило всю отрасль – и теперь целый хор голосов в сообществе разработчиков ИИ говорит о том, что, если мы не будем достаточно осторожны, мир кончит *именно* так. По крайней мере на сегодняшний момент человечество проигрывает эту игру.

* * *

Эта книга посвящена машинному обучению и человеческим ценностям. В ней рассказывается о системах, которые обучаются с помощью данных, а не жестко запрограммированы, и о том, как и чему именно мы пытаемся их научить.

Область машинного обучения включает в себя три крупных сферы: при неконтролируемом обучении (обучении без учителя) машине просто дают множество данных и – как в

случае с системой Word2Vec – просят найти в них смысл, определить схемы, закономерности, полезные способы редукции, представления или визуализации данных. При контролируемом обучении системе дают ряд упорядоченных по категориям или размеченных примеров – как в случае с условно-досрочно освобожденными, часть из которых снова попала под арест, а часть осталась на свободе, – и приказывают сделать прогноз о новых примерах, которые система не видела при обучении, или по тем, ответы по которым и вовсе еще неизвестны. При обучении с подкреплением систему помещают в обстановку, где существуют награды и наказания, такую, как лодочная гонка с заправками и повреждениями – и приказывают найти лучший способ свести к минимуму наказания и максимизировать награды.

По всем трем направлениям все чаще кажется, что эти математические и компьютерные модели так или иначе переворачивают мир. Отличаясь по сложности – от систем, которые ведут книги бухгалтерского учета, до тех, которые можно с большой долей уверенности назвать *искусственным интеллектом*, – они неуклонно замещают как человеческое суждение, так и программное обеспечение более традиционного вида.

Этот процесс происходит не только в сфере технологий, не только в торговле, но и в областях, где есть этическая и моральная нагрузка. Законы штатов и федеральные законы дают все больше полномочий на использование программ-

ного обеспечения для оценки рисков при освобождении под поручительство и условно-досрочном освобождении. Легковые и грузовые автомобили на шоссе и на соседних улицах все чаще управляют сами собой. Мы больше не можем предполагать, что наше заявление на получение ипотеки, резюме или результаты анализов увидит хоть один человек до того, как решение будет принято. Будто бóльшая часть человечества в начале XXI века поглощена задачей постепенного перевода мира на автопилот – как в переносном смысле, так и буквально.

В последние годы сигналы тревоги раздаются сразу в двух абсолютно независимых друг от друга сообществах. Первое – это люди, обратившие внимание на современные этические риски, связанные с технологиями. Если система распознавания лиц очень неточна в работе с людьми определенной расы или пола, но прекрасно справляется с другими, или если кому-то отказали в освобождении под залог из-за того, что статистическую модель никогда не проверяли, а в зале суда не было ни одного человека – ни судьи, ни прокурора, ни защитника, – понятно, что это проблема. Такие задачи не могут быть решены в рамках традиционных отраслей науки, нужен диалог между специалистами по теории вычислительных машин и систем, социологами, юристами, политологами, этиками. Диалог этот начался в спешке.

Второе сообщество – это те, кто обеспокоен будущими угрозами, которые поджидают нас по мере того, как наши

системы наращивают возможности по быстрому и гибкому принятию решений и в Интернете, и в реальном мире. В последнее десятилетие мы, бесспорно, могли наблюдать самый головокружительный, скачкообразный и вызывающий беспокойство прогресс за всю историю машинного обучения и даже искусственного интеллекта. Существует негласное соглашение, разрушившее своего рода табу: исследователям ИИ больше не запрещено обсуждать вопросы безопасности. На самом деле, такая озабоченность в последние пять лет вышла из кулуаров, став одной из центральных в отрасли.

Несмотря на споры о том, следует ли отдавать приоритет непосредственным или долговременным проблемам, эти два сообщества едины в своих более масштабных целях.

Поскольку системы машинного обучения не только распространяются, но и становятся все более мощными, мы чаще оказываемся в положении волшебника-недоучки: обращаемся к существующей отдельно от нас, но полностью послушной силе, даем ей ряд инструкций, а потом носимся как сумасшедшие, пытаясь остановить ее, поняв, что наши инструкции неточны или неполны, хотя, как это ни ужасно, получаем именно то, о чем просили.

Как избавиться от столь катастрофического расхождения целей и результатов, как убедиться, что модели учитывают наши нормы и ценности, понимают, что мы имеем в виду и что намереваемся сделать и, самое главное, чего мы на самом деле хотим? Этот вопрос стал одной из самых главных

и срочных проблем в области информатики. У него есть название – *проблема выравнивания*.

Реакция на этот сигнал тревоги – как на то, что передний край науки как никогда близко подошел к разработке так называемого «общего» интеллекта, так и на то, что существующие в реальном мире системы машинного обучения все больше и больше касаются этически неоднозначных аспектов личной и общественной жизни – была неожиданной и энергичной. Самые разные люди призывают действовать, выходя за рамки традиционных дисциплинарных направлений. Некоммерческие организации, научно-исследовательские центры и институты упрочивают свое положение. Лидеры как в бизнесе, так и в науке начинают открыто высказываться, предостерегая и выражая свою озабоченность, а также перераспределяют финансирование исследований в соответствии с новыми опасениями. На программы постдипломного обучения зачислено первое поколение студентов, работы которых направлены непосредственно на этическую составляющую и безопасность машинного обучения. Первые отклики на проблему выравнивания уже появились.

Конец ознакомительного фрагмента.