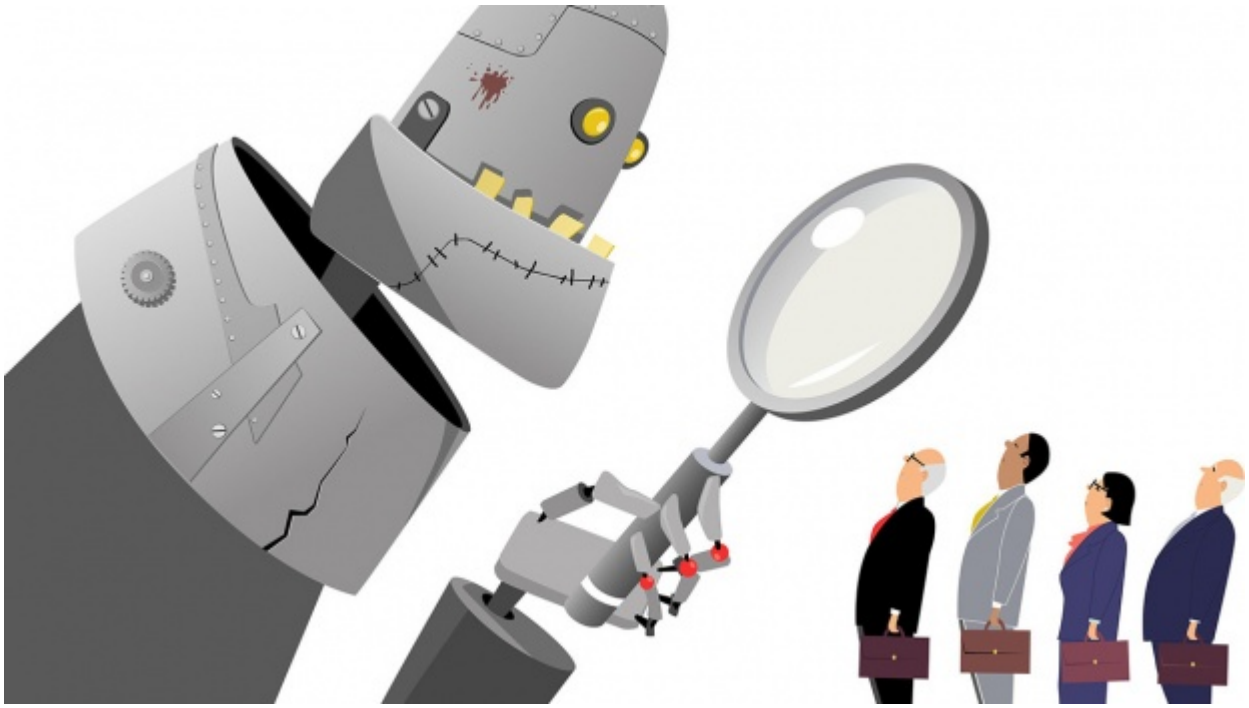
Почему искусственный интеллект предвзят?

Юлия Каленкова

https://te-st.org/2019/11/29/why-is-artificial-intelligence-biased/

Статья обновлена 06 июня 2023



Что такое «предвзятость искусственного интеллекта» (AI bias)? С чем связано возникновение этого явления, и как с ним бороться? На эти вопросы попытались ответить эксперты научно-популярного журнала Harvard Business Review.

Термин AI bias можно перевести как «необъективность ИИ» или «пристрастность ИИ». Причина столь высокого интереса к AI bias объясняется тем, что результаты внедрения технологий ИИ в ряде случаев задевают основные ценности современного общества. Они проявляются в нарушении таких важных принципов, как расовое и гендерное равенства. IT-специалисты рекомендуют проверять данные, поступающие в системы ИИ, чтобы они не содержали «исторического предубеждения против определенных групп».

В статье «Вот почему возникают ИИ-привязанности и почему с ними сложно бороться» выделяются три момента, способствующие возникновению AI bias.

- Постановка задачи (Framing the problem). Проблема состоит в том, что методами машинного обучения обычно хочется опередить нечто, не имеющее строгого определения. Скажем, банк хочет определить кредитные качества заемщика, но это весьма размытое понятие и результат работы модели будет зависеть от того, как разработчики, в силу своих личных представлений, смогут это качество формализовать.
- Сбор данных для обучения (Collecting the data). На данном этапе может быть два

источника предвзятости: данные могут быть не репрезентативны или же могут содержать предрассудки. Известный прецедент, когда система лучше различала светлокожих по сравнению с темнокожими, был связан с тем, что в исходных данных светлокожих было больше. А не менее известная ошибка в автоматизированных рекрутинговых службах, которые отдавали предпочтения мужской половине, была связаны с тем, что они были обучены на данных, страдающих мужским шовинизмом.

- Подготовка данных (Preparing the data). Когнитивная предвзятость может просочиться при выборе тех атрибутов, которые алгоритм будет использовать при оценке заемщика или кандидата на работу. Никто не может дать гарантии объективности избранного набора атрибутов.

Системы машинного обучения игнорируют показатели, которые не могут точно предсказать результат (среди данных, предложенных для обучения). А значит, необходима дополнительная проверка алгоритмов. Она должна найти ошибки, заложенные человеком, те, что остались незамеченными или непроверенными.

> Алгоритм обычно предсказывает будущее точно, но не говорит, ни чем будет вызвано событие, ни почему. Алгоритм может прочитать все статьи New York Times и сказать, что из них будут обсуждать в «Твиттере», но не объяснит желание людей поделиться этой информацией. Алгоритм может показать, какие сотрудники, скорее всего, многого добьются, но не сообщит вам, в силу каких своих качеств. Понять эти два недостатка алгоритмов — значит, сделать первый шаг к тому, чтобы лучше ими управлять.

Главное, что мы должны сделать, – это ускорить прогресс, который наблюдается в работе над ошибками в искусственном интеллекте. Один из наиболее сложных шагов в этом направлении – понимание и оценка «достоверности». Исследователи разработали технические способы определения достоверности.

Например, требование, чтобы модели имели равную прогностическую ценность среди групп, или требование, чтобы модели имели одинаковое число ложноположительных и отрицательных распознаваний среди групп. Однако это приводит к сложной задаче – различные требования по достоверности обычно не могут быть удовлетворены одновременно.

«Достоверность противоречий» и другие решения проблемы AI bias

Решением проблемы AI bias, по мнению экспертов Harvard Business Review, может служить заблаговременная обработка и анализ данных или внедрение значений достоверности непосредственно в процесс обучения системы. Одной из перспективных технологий является «достоверность противоречий», в рамках которой проверяется, что решения, принимаемые алгоритмами, будут оставаться неизменными в противоположной плоскости, где изменены такие важные переменные, как раса, пол и сексуальная ориентация.

> Еще по теме: Этика и математика: зачем кодировать мораль

Сильвия Кьяппа из компании DeepMind разработала особенный подход к «достоверности противоречий». Он позволяет решать сложные задачи, когда некоторые ветви алгоритма, в которых значимые переменные влияют на результат, считаются достоверными, тогда как

другие факторы влияния признаются фиктивными.

Например, эта модель может помочь убедиться в том, что зачисление абитуриента на тот или иной факультет не зависело от его пола, при этом сохранив влияние полового признака на общее зачисление студентов в университет (допустим, абитуриентки подают заявки на факультеты с более высоким конкурсом).

Но другие задачи требуют чего-то большего, чем просто техническое решение. Например, как определить, что система дает достаточно достоверные результаты и готова к использованию? И в каких случаях вообще следует разрешать автоматизированное принятие решений? Такие вопросы требуют многопрофильного подхода, в том числе специалистов по этике, специалистов в области общественных наук и гуманитарных дисциплинах.

Как бороться с погрешностями искусственного интеллекта

- Лидеры бизнеса и общества должны быть в курсе трендов ИИ, поскольку многие организации могут дать источники дополнительной информации, например, годовые отчеты, данные некоммерческого сектора.
- Внедряя искусственный интеллект, нужно стараться минимизировать систематические ошибки. Например, используя уже имеющийся набор технических средств и методик. Так, подразделение Google AI опубликовало список рекомендуемых практик, а компания IBM в своем фреймворке «Fairness 360» собрала наиболее распространенные инструменты.
- Должна вестись параллельная работа алгоритмов и человека. Важно отметить, что когда мы действительно нашли ошибку, недостаточно просто изменить алгоритм. Необходимо улучшать процессы деятельности человека, которые лежат в основе этих алгоритмов.
- Нужно понять, каким образом люди и машины могут работать сообща для выявления ошибок. Понимание степени уверенности, с которой алгоритмы дают рекомендации, позволит определить, какую свободу принятия решений им следует давать.
- Не менее чем многопрофильный подход в исследовании систематических ошибок, важна конфиденциальность данных. Этическая составляющая будет способствовать коммуникации между человеком и умными машинами.
- Вовлечение заинтересованных сообществ сделает развитие ИИ всесторонним и эффективным. По такому принципу работает AI4ALL – некоммерческая организация, которая путем обучения и менторства готовит многопрофильных специалистов в области искусственного интеллекта в недостаточно представленных социальных группах.

У искусственного интеллекта много перспектив – как в социальной сфере, так и в бизнесе. Но их получится реализовать, только если люди будут доверять алгоритмам.

Еще по теме

- Социализация искусственного интеллекта: будут ли роботы чувствовать?
- Искусственный интеллект и человеческая этика: мир будущего
- Есть ли у искусственного интеллекта права и обязанности?