DOI 10.58351/2949-2041.2026.31.2.010

Жижкин Михаил Александрович, студент
Московский государственный технический университет
имени Н. Э. Баумана

БОРЬБА С АЛГОРИТМИЧЕСКИМИ ПРЕДВЗЯТОСТЯМИ

Аннотация. В работе кратко рассматриваются причины появления предвзятостей в алгоритмах машинного обучения, их социальные и правовые последствия, а также современные подходы к выявлению и снижению алгоритмической дискриминации (качество и разнообразие данных, прозрачность алгоритмов и междисциплинарная экспертиза).

Ключевые слова: Алгоритмическая предвзятость; данные обучения; скрытая дискриминация; Learning Fair Representations; прозрачность алгоритмов; этика ИИ.

Введение

Алгоритмическая предвзятость – это систематическая и повторяющаяся склонность модели выдавать несправедливые решения по отношению к определённым группам людей. Источниками такой предвзятости выступают прежде всего реальные данные, на которых учатся модели: если в прошлом отдельные категории населения подвергались дискриминации или имели ограниченный доступ к ресурсам, соответствующие паттерны могут быть «усвоены» алгоритмом и воспроизведены автоматически. Законодательная и международно-правовая база подчёркивает недопустимость дискриминации и равенство перед законом (см. Международный пакт о гражданских и политических правах), однако правовые нормы сами по себе не решают всех технических и организационных проблем, связанных с предвзятостью ИИ.

1. Причины появления предвзятостей в алгоритмах машинного обучения

Алгоритмическая предвзятость (англ. *algorithmic bias*) – это систематическая и повторяющаяся вредоносная склонность компьютерной системы или социотехнической системы, приводящая к появлению «несправедливых» результатов, например, «привилегированию» одной категории над другой вопреки задуманной функции алгоритма [3, pp. 125 – 142].

Модели МО (машинного обучения) обучаются на данных, собранных из реального мира. В случае, если в предшествующем конкретные категории людей сталкивались с дискриминацией либо ограничениями, алгоритмы имеют все шансы воспринимать эти паттерны. К примеру, если в определённых профессиях преимущественно были заняты представители сильного пола, алгоритм способен воспринимать девушек в данных ролях в виде исключения, что приводит к понижению их шансов на трудоустройство в определенном месте. При разработке алгоритмов важно учитывать баланс и разнообразие данных. Если модель обучается на несбалансированных данных, она может неправильно интерпретировать закономерности. К примеру, метод, предназначенный с целью моделирования преступности, способен ошибочно рассматривать конкретные группы населения наиболее предрасположенными к преступлениям (например, предвзятое отношение правоохранительных органов к определенным этническим группам, как в США). Разработчики алгоритмов часто представляют собой однородную группу с похожими взглядами и опытом. Это может привести к созданию систем, которые не учитывают разнообразие и потребности различных групп населения. Например, системы распознавания лиц, обученные преимущественно на изображениях определённых этнических групп, могут плохо работать с лицами представителей других групп. Предвзятость алгоритмов ИИ может быть частично ограничена стандартами и этическими принципами, закрепленными в программных документах. Но одно лишь нормативное правовое регулирование неспособно устранить эту проблему. Во-первых, разнообразие принципов, рамок, руководств и стандартов

создает возможность лавировать между ними, чтобы оправдать существующие алгоритмы «как есть», не внося в них изменения, а также создает проблему несопоставимости многих стандартов. Во-вторых, как было отмечено выше, предвзятость алгоритмов может в значительной степени порождаться предвзятыми решениями в реальной жизни, которые алгоритм «усваивает» во время обучения на реальных данных. Разработка механизмов выявления и корректировки этого «усвоения», таким образом, становится ключевым элементом в преодолении проблем неэтичности алгоритмов ИИ [1, с. 118–126].

Предвзятость в алгоритмах – это не просто техническая проблема, а серьёзный социальный вызов. Для её решения необходимо:

1. Гарантировать, чтобы данные, в которых учатся алгоритмы, отображали многообразие реального общества.
2. Создатели обязаны являться открытыми в отношении методов и информации, применяемых с целью изучения моделей, для того чтобы общество имело возможность дать оценку и равным образом опровергнуть выводы, принимаемые алгоритмами.
3. Введение профессионалов из различных областей, в том числе социологов, специалистов по психологии и правозащитников, в процесс разработки и оценки алгоритмов.

Восприятие и предотвращение предвзятости в алгоритмах – это этап к формированию наиболее объективного сообщества, где технологические процессы предназначаются на благо всех без исключения, а не только отдельных групп. Следовательно, основная причина – это предвзятые обучающие данные.

2. Примеры дискриминации в ИИ

Дискриминация человеческого капитала тесно связана с принадлежностью человека к определенной группе, однако, как известно, ни один вид группового членства не имеет права на проявление какой-либо предвзятости по отношению к другому виду. Правовое положение о дискриминации отражено в статье 26 Международного пакта о гражданских и политических правах: «Все люди равны перед законом и имеют право на равную защиту закона без любых проявлений дискриминации. Справедливый закон запрещает любую дискриминацию и обеспечивает гарантию равноправия для всех, также закон должен обеспечивать эффективную защиту от дискриминации по любому основанию, таким как раса, цвет кожи, пол, язык, религия, политические или иные взгляды, национальное или социальное происхождение, имущество, рождение или другой статус». Казалось бы, что все вопросы, связанные с дискриминацией, изучены и обрели правовой статус. Однако на сегодняшний момент возникает новое проявление предвзятости, которое весьма трудно предугадать и предусмотреть. Эта новая волна дискриминации связана с распространением и проникновением во все сферы жизни ИИ. ИИ действительно начинает технологическую революцию, и, хотя ему еще предстоит «захватить» мир, есть более насущная проблема, с которой мы уже столкнулись, – предвзятость ИИ. Предвзятость ИИ – это основное предубеждение в данных, которые используются для создания алгоритмов ИИ, что в конечном итоге может привести к дискриминации и другим последствиям. Приведем простой пример. Представим, что мы хотим создать алгоритм, который решает, будет ли абитуриент принят в университет или нет, и одним из наших входных данных будет географическое местоположение абитуриента. Гипотетически, если бы местоположение человека сильно коррелировало с этнической принадлежностью, то наш алгоритм косвенно отдавал бы предпочтение определенным этническим группам перед другими. Это пример предвзятости в ИИ [2, с. 179].

Как мы уже знаем, основная причина – это предвзятые обучающие данные. Если данные, на которых обучает модель, уже имеет признаки перекоса к определенным группам лиц (например, определённая группа чаще получала отказы в выдаче кредита в прошлом), то ИИ просто «научится» повторять эту несправедливость. Вторая причина – выбор признаков, который может быть косвенно связан с сентиментальными характеристиками, даже если напрямую они никак не учитываются при скрытой дискриминации. Самое опасное – это

скрытая дискриминация т.е. когда чувствительный признак не отображается в данных, на которых модель обучалась, но алгоритм угадывает его путём определенных логических размышлений в ходе обучения. Чисто технически, определенный уровень дохода, место проживания, тип занятости и т.п. может косвенно отразить расу или пол человека. Из-за этого хочется поднять такой вопрос: «ИИ точно может быть справедливым?»

3. Методы выявления и устранения алгоритмических предвзятостей

Алгоритмическая предвзятость – сложная проблема, которая требует внимательного отношения разработчиков к ней, проявляется в виде определённого дисбаланса или перевеса в ту или иную группу пользователей. Его суть в том, чтобы утратить любую информацию, которая может идентифицировать принадлежность человека к защищенной подгруппе.

Данная модель уже была предложена некоторыми учёными за рубежом. Она называется *Learning Fair Representations* (далее LFR) – подхода, в котором при обучении модели создаётся промежуточное представление данных, утрачивающее любую информацию о принадлежности к защищённой группе, но сохраняющее достаточную полезность для решения задачи. В этой статье мы не будем вдаваться в технические подробности, но одно можно сказать точно, на данный момент это одна из самых продвинутых идей по решению сложившийся ситуации. Метод LFR был впервые предложен Р. Земелем и соавторами в работе *"Learning Fair Representations"* (ICML, 2013) [Zemel et al., 2013].

Цель LFR – трансформировать исходные данные в новое представление, которое остаётся полезным для основной задачи (например, классификации), но скрывает информацию о чувствительном (защищённом) признаке так, чтобы алгоритм не мог «узнать» принадлежность к чувствительной группе.

Это достигается путём обучения стохастического отображения входов (x) в представление (z), контролируя одновременно:

- точность задачи,
- «справедливость» (fairness),
- способность к реконструкции исходных данных.

Таким образом, модель создаёт промежуточное представление данных в виде вероятностного распределения над набором прототипов – типичных "синтетических" примеров. Каждый реальный объект отображается как вектор вероятностей принадлежности к этим прототипам. Алгоритм стремится сделать это отображение таким, чтобы оно не позволяло предсказать чувствительный признак, но при этом оставалось полезным для основной задачи (например, классификации дохода). Однако, следует отметить, что достижение полной нейтральности к текущему признаку ведет к снижению точности модели. Поэтому тут позволяется настраивать баланс в зависимости от конкретной задачи.

Представьте, вы создаете модель, для того чтобы прогнозировать, получит ли человек кредит. Обыкновенная модель может косвенно обучиться дискриминировать, предположим, женщин через косвенные признаки, например, по количеству детей либо профессии. *Метод LFR* позволяет создать промежуточное представление z, в котором такая корреляция с полом отсутствует, в результате женщина с тем же профилем, что и мужчина, будет получать одинаковое решение по кредиту. Исходя из этого, метод LFR дает возможность встроить принципы справедливости непосредственно в процесс обучения модели, обеспечивая сбалансированные решения в отсутствии очевидной либо утаенной дискриминации.

Заключение

Предвзятость в алгоритмах машинного обучения коренится прежде всего в данных и в организационных практиках разработки. Для снижения рисков необходим комплекс мер: обеспечение репрезентативности и качества данных, исключение и контроль proxy-признаков, внедрение технических приёмов (например, LFR и другие методы справедливого обучения), повышение прозрачности моделей и привлечение специалистов из социальных наук и правозащитных сфер. Законодательные нормы важны для установки стандартов, но их

эффективность возрастёт только при сочетании с техническими решениями и корпоративной ответственностью. В долгосрочной перспективе достижение справедливых алгоритмов требует осознанного выбора приоритета – компромисса между точностью и этическими требованиями – и постоянного мониторинга моделей в реальной эксплуатации

Список литературы:

1. Смирнова А. И. Предвзятость как проблема алгоритмов ИИ: этические аспекты// Философия и общество. – 2023. – № 3. – С. 118–126.

2. Каштанова Е. В., Лобачева А. С. Проблемы предвзятости и дискриминации человеческого капитала в системах искусственного интеллекта// Вестник университета. – 2024. – №3. С. 176 – 185.

3. Hardebolle, Cécile, Héder, Mihály, Ramachandran, Vivek. Engineering ethics education and artificial intelligence/ The Routledge International Handbook of Engineering Ethics Education. – Lausanne, 2024. – P. 125 – 142