УДК 004.89

ИНСТРУМЕНТЫ АНАЛИЗА ПРЕДВЗЯТОСТИ И СПРАВЕДЛИВОСТИ В ПРЕДОБУЧЕННЫХ МОДЕЛЯХ

Голошумов И.Е., Оператор научной роты, III курс
Военная академия связи имени Маршала Советского Союза С. М. Буденного,
г. Санкт-Петербург

Внедрение предобученных моделей машинного обучения в социально значимые и регулируемые области – от оценки кандидатов при приёме на работу до рекомендаций в кредитных и медицинских системах – требует не только высокой точности, но и справедливости и непредвзятости. Однако многочисленные исследования, включая работы MIT Media Lab, Google AI и Microsoft Research, демонстрируют, что даже самые современные и масштабные модели, включая трансформеры, визуальные классификаторы и мультимодальные системы, демонстрируют систематическую предвзятость, унаследованную от данных, архитектурных ограничений или стратегий обучения. Это может выражаться в дискриминации по полу, расе, возрасту, географическому признаку, акценту, инвалидности или даже социальной роли [1].

В условиях, когда вывод модели используется в реальных процедурах принятия решений, такие искажения могут приводить к социальным, правовым и этическим нарушениям. Именно поэтому разработка и внедрение инструментов анализа предвзятости и справедливости становятся ключевым направлением инженерии доверенного искусственного интеллекта.

Предвзятость в моделях может проявляться как в явной форме – например, если модель систематически занижает рейтинг резюме женщин или людей определённой национальности, – так и в неявной форме, когда дискриминация возникает через опосредованные признаки (например, адрес проживания, стиль речи, частотность определённых слов). Особенно сложно выявлять структурную предвзятость, при которой сама постановка задачи, формулировка целей или выбор целевых меток уже содержит социокультурные искажения. В таких случаях сама модель может быть статистически точной, но при этом несправедливой по отношению к отдельным подгруппам.

Инструменты анализа предвзятости можно условно разделить на диагностические, метрические и визуализирующие, при этом большинство современных фреймворков сочетают все три направления.

Одним из центральных классов таких инструментов являются метрики справедливости (fairness metrics), позволяющие количественно оценить

степень неравномерности поведения модели относительно разных групп пользователей или объектов.

Наиболее часто используются:

- Demographic parity – модель должна давать одинаковую вероятность положительного исхода для всех демографических групп.
- Equalized odds – вероятность положительного предсказания должна быть одинаковой при одинаковом истинном классе.
- Predictive parity – позитивные предсказания должны быть одинаково точны между группами.
- Treatment equality – соотношение ложноотрицательных и ложноположительных должно быть схожим между группами.

Каждая из этих метрик отражает разную концепцию справедливости (процедурную, результативную, групповую или индивидуальную), и зачастую они взаимно несовместимы: достижение одной может нарушать другую. Поэтому инструменты анализа должны не просто рассчитывать метрики, но и визуализировать компромиссы, позволяя выбирать стратегию в зависимости от контекста.

Одним из наиболее популярных фреймворков является IBM AI Fairness 360 (AIF360), предоставляющий более 70 метрик и 10 алгоритмов коррекции моделей. Он поддерживает как бинарную классификацию, так и регрессию, и может применяться как к моделям «в чёрном ящике», так и к полным пайплайнам с контролем данных. С помощью AIF360 можно не только анализировать, но и модифицировать модель для повышения её справедливости – например, путём перекодирования входных данных, изменения весов примеров или коррекции выводов.

Другой важный инструмент – Fairlearn (от Microsoft Research), который позволяет строить модели с учётом ограничений на справедливость. В отличие от AIF360, он фокусируется на оптимизации справедливости как части целевой функции, позволяя напрямую задавать желаемые метрики справедливости при сохранении максимально возможной точности. Это особенно важно в задачах, где требуется баланс между эффективностью и социальной допустимостью модели, например, при отборе на обучение, распределении социальной помощи или автоматической диагностике [2].

Для языковых моделей существует специализированный класс тестов, таких как StereoSet, WinoBias, CrowS-Pairs, которые позволяют выявить лингвистические паттерны предвзятости. Например, если модель при дополнении фразы "The nurse walked into the room and saw the..." чаще продолжает её как "doctor" при мужском контексте и "patient" при женском, это свидетельствует о стереотипном распределении вероятностей. Современные языковые модели

часто демонстрируют предвзятость, особенно если обучены на открытых интернет-корпусах без фильтрации [3].

Визуальные модели анализируются с помощью наборов данных, таких как Gender Shades или FairFace, в которых представлены сбалансированные изображения по полу и расе. Оказалось, что даже коммерческие модели распознавания лиц имеют существенно меньшую точность на тёмнокожих женщинах по сравнению с белыми мужчинами. Это вызвало международный резонанс и стало толчком к пересмотру стандартов тестирования в компьютерном зрении. Современные инструменты визуального анализа – например, What-If Tool, Facets или TCAV (Testing with Concept Activation Vectors) – позволяют оценить чувствительность модели к интерпретируемым признакам и выявить, какие именно компоненты входа влияют на её поведение [4].

Особую роль играют инструменты каузального анализа, позволяющие отличить корреляционные зависимости от причинных. Например, если модель различает классы не по признаку заболевания, а по качеству сканера (который коррелирует с типом больницы и, соответственно, социоэкономическим статусом пациентов), это указывает на предвзятость, которая не выявляется обычными метриками. Каузационный анализ позволяет проводить интервенции на уровне признаков и оценивать устойчивость вывода модели к условным модификациям. Этот подход постепенно внедряется в библиотеки, такие как DoWhy, EconML и CausalML [5].

Справедливость может рассматриваться и как элемент нормативной совместимости. Например, в соответствии с Законом о цифровых услугах (Digital Services Act) и готовящимся AI Act в ЕС, организации, использующие модели ИИ в чувствительных областях, обязаны оценивать и документировать риски предвзятости, а также предоставлять объяснение отказа в обслуживании или решении. Это означает, что инструменты анализа предвзятости должны быть интегрированы в производственные пайплайны, обеспечивать валидацию моделей на этапах верификации и использоваться для формирования документации [6].

Таким образом, инструменты анализа предвзятости и справедливости в предобученных моделях представляют собой неотъемлемый элемент инфраструктуры доверенного ИИ. Они позволяют не только диагностировать и объяснять ошибки, но и строить системы, учитывающие ценности, нормативы и ожидания общества.

В условиях растущей автоматизации принятия решений вопрос справедливости становится не этическим дополнением, а технологическим требованием, без выполнения которого ни одна модель не может быть по-настоящему надёжной, транспарентной и социально легитимной.

Список литературы:

1. Barocas, S. Fairness and Machine Learning : Limitations and Opportunities / S. Barocas, M. Hardt, A. Narayanan. – [S. l.] : fairmlbook.org, 2019. – URL: https://fairmlbook.org/ (дата обращения: 09.06.2025).

2. Dwork, C. Fairness through Awareness / C. Dwork, M. Hardt, T. Pitassi [et al.] // Proceedings of the 3rd Innovations in Theoretical Computer Science Conference. – New York : ACM, 2012. – P. 214–226. – DOI 10.1145/2090236.2090255.

3. Hardt, M. Equality of Opportunity in Supervised Learning / M. Hardt, E. Price, N. Srebro // Advances in Neural Information Processing Systems. – 2016. – Vol. 29.

4. Chouldechova, A. Fair Prediction with Disparate Impact : A Study of Bias in Recidivism Prediction Instruments / A. Chouldechova // Big Data. – 2017. – Vol. 5, № 2. – P. 153–163. – DOI 10.1089/big.2016.0047.

5. Kleinberg, J. Inherent Trade-Offs in the Fair Determination of Risk Scores / J. Kleinberg, S. Mullainathan, M. Raghavan // Proceedings of Innovations in Theoretical Computer Science. – Schloss Dagstuhl : LIPIcs, 2017. – Vol. 67. – P. 43:1–43:23. – DOI 10.4230/LIPIcs.ITCS.2017.43.

6. Saleiro, P. Aequitas : A Bias and Fairness Audit Toolkit / P. Saleiro, B. Kuester, A. Stevens [et al.] // arXiv. – 2018. – arXiv:1811.05577.