ФЕДЕРАЛЬНОЕ АГЕНТСТВО

ПО ТЕХНИЧЕСКОМУ РЕГУЛИРОВАНИЮ И МЕТРОЛОГИИ

ПРЕДВАРИТЕЛЬНЫЙ НАЦИОНАЛЬНЫЙ СТАНДАРТ РОССИЙСКОЙ ФЕДЕРАЦИИ

ПНСТ 839— 2023 (ISO/IEC TR 24027:2021)

ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ

Смещенность в системах искусственного интеллекта и при принятии решений с помощью искусственного интеллекта

(ISO/IEC TR 24027:2021, Information technology — Artificial intelligence (AI) — Bias in AI systems and AI aided decision making, MOD)

Издание официальное

Москва
Российский институт стандартизации
2023

Предисловие

1 ПОДГОТОВЛЕН Федеральным государственным автономным образовательным учреждением высшего образования «Национальный исследовательский университет «Высшая школа экономики» (НИУ ВШЭ) на основе собственного перевода на русский язык англоязычной версии документа, указанного в пункте 4

2 ВНЕСЕН Техническим комитетом по стандартизации ТК 164 «Искусственный интеллект»

3 УТВЕРЖДЕН И ВВЕДЕН В ДЕЙСТВИЕ Приказом Федерального агентства по техническому регулированию и метрологии от 10 ноября 2023 г. № 51-пнст

4 Настоящий стандарт является модифицированным по отношению к международному документу ISO/IEC TR 24027:2021 «Информационные технологии. Искусственный интеллект (ИИ). Смещенность в системах ИИ и при принятии решений с помощью ИИ» (ISO/IEC TR 24027:2021 «Information technology — Artificial intelligence (AI) — Bias in AI systems and AI aided decision making», MOD). При этом дополнительные слова (фразы, показатели, ссылки), включенные в текст стандарта для учета особенностей российской национальной стандартизации, выделены полужирным курсивом, а объяснения причин их включения приведены в сносках.

Наименование настоящего стандарта изменено относительно наименования указанного международного документа для приведения в соответствие с ГОСТ Р 1.5—2012 (пункт 3.5).

Сведения о соответствии ссылочных национальных стандартов международным стандартам, использованным в качестве ссылочных в примененном международном документе, приведены в дополнительном приложении ДА

Правила применения настоящего стандарта и проведения его мониторинга установлены в ГОСТ Р 1.16—2011 (разделы 5 и 6).

Федеральное агентство по техническому регулированию и метрологии собирает сведения о практическом применении настоящего стандарта. Данные сведения, а также замечания и предложения по содержанию стандарта можно направить не позднее чем за 4 мес до истечения срока его действия разработчику настоящего стандарта по адресу: info@tc164.ru и/или в Федеральное агентство по техническому регулированию и метрологии по адресу: 123112 Москва, Пресненская набережная, д. 10, стр. 2.

В случае отмены настоящего стандарта соответствующая информация будет опубликована в ежемесячном информационном указателе «Национальные стандарты» и также будет размещена на официальном сайте Федерального агентства по техническому регулированию и метрологии в сети Интернет (www.rst.gov.ru)

Содержание

Введение

Смещенность (предвзятость) в системах искусственного интеллекта (ИИ) может проявляться по-разному. Системы ИИ, которые изучают закономерности на основе данных, потенциально могут отражать существующие в обществе предубеждения в отношении каких-либо групп. Хотя некоторая смещенность необходима для решения задач системы ИИ (т.е. требуемая смещенность — desired bias), может существовать смещенность, которая не предусмотрена задачами и, таким образом, представляет собой нежелательную смещенность (unwanted bias) в системе ИИ.

Смещенность в системах ИИ может возникнуть в результате структурных недостатков при проектировании системы, из-за когнитивных предубеждений заинтересованных лиц или быть присущей наборам данных, используемым для обучения моделей. Это означает, что системы ИИ могут надолго закрепить или усилить существующие предубеждения или создать новые предубеждения.

Разработка систем ИИ с результатами, свободными от нежелательной смещенности, является сложной задачей. Функциональное поведение систем ИИ является сложным и может быть трудным для понимания, но исключить нежелательную смещенность возможно. Многие виды деятельности по разработке и внедрению систем ИИ предоставляют возможности для выявления и устранения нежелательных предубеждений, чтобы заинтересованные стороны могли воспользоваться преимуществами систем ИИ в соответствии со своими целями.

Смещенность в системах ИИ является областью активных исследований. В настоящем стандарте изложены современные передовые методы выявления и устранения смещенности в системах ИИ или при принятии решений с помощью ИИ, независимо от источника. Настоящий стандарт охватывает такие темы, как:

- обзор смещенности (5.2) и справедливости (5.3);
- потенциальные источники нежелательной смещенности и термины для определения характера потенциальной смещенности (раздел 6);
- оценка смещенности и справедливости (раздел 7) с помощью метрик;
- устранение нежелательной смещенности с помощью различных стратегий (раздел 8).

ПРЕДВАРИТЕЛЬНЫЙ НАЦИОНАЛЬНЫЙ СТАНДАРТ РОССИЙСКОЙ ФЕДЕРАЦИИ

ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ

Смещенность в системах искусственного интеллекта и при принятии решений с помощью искусственного интеллекта

Artificial intelligence. Bias in artificial intelligence systems and artificial intelligence aided decision making

**Срок действия — с 2024—01—01
до 2027—01—01**

1 Область применения

В настоящем стандарте рассматриваются вопросы смещенности в системах ИИ, особенно в отношении принятия решений с помощью ИИ. Описаны методы измерения и методы оценки смещенности с целью устранения и обработки уязвимостей, связанных со смещенностью. Рассматриваются все фазы жизненного цикла системы ИИ, включая сбор данных, обучение, непрерывное обучение, проектирование, тестирование, оценку и использование, но обзор не ограничивается рассмотрением только этих процессов.

2 Нормативные ссылки

В настоящем стандарте использована нормативная ссылка на следующий стандарт:

ПНСТ 838—2023 (ИСО/МЭК 23053:2022) Искусственный интеллект. Структура описания систем искусственного интеллекта, использующих машинное обучение.

Примечание — При пользовании настоящим стандартом целесообразно проверить действие ссылочных стандартов в информационной системе общего пользования — на официальном сайте Федерального агентства по техническому регулированию и метрологии в сети Интернет или по ежегодному информационному указателю «Национальные стандарты», который опубликован по состоянию на 1 января текущего года, и по выпускам ежемесячного информационного указателя «Национальные стандарты» за текущий год. Если заменен ссылочный стандарт, на который дана недатированная ссылка, то рекомендуется использовать действующую версию этого стандарта с учетом всех внесенных в данную версию изменений. Если заменен ссылочный стандарт, на который дана датированная ссылка, то рекомендуется использовать версию этого стандарта с указанным выше годом утверждения (принятия). Если после утверждения настоящего стандарта в ссылочный стандарт, на который дана датированная ссылка, внесено изменение, затрагивающее положение, на которое дана ссылка, то это положение рекомендуется применять без учета данного изменения. Если ссылочный стандарт отменен без замены, то положение, в котором дана ссылка на него, рекомендуется применять в части, не затрагивающей эту ссылку.

3 Термины и определения

В настоящем стандарте применены термины по ***[1]***[1)] и ***ПНСТ 838—2023***, а также следующие термины с соответствующими определениями:

[1)] Нормативная ссылка на международный стандарт ИСО/МЭК 22989 заменена на справочную ссылку из-за отсутствия гармонизированного с ним национального стандарта и его перевода на русский язык в Федеральном информационном фонде стандартов.

Издание официальное

3.1 Искусственный интеллект

3.1.1 **метод максимального правдоподобия** (maximum likelihood estimator): Оценка, определяющая значение параметра, при котором функция правдоподобия достигает или приближается к своему наибольшему значению.

Примечание — Метод максимального правдоподобия является хорошо известным подходом для получения оценок параметров, когда задано распределение (например, нормальное, гамма, Вейбулла и т. д.). Эти оценки обладают требуемыми статистическими свойствами (например, инвариантностью при монотонном преобразовании) и во многих ситуациях являются предпочтительным методом оценки. В случаях, когда оценка максимального правдоподобия является смещенной, иногда проводится простая коррекция смещения ***[2]***.

3.1.2 **системы, основанные на правилах** (rule-based systems): Система, основанная на знаниях, которая делает выводы, применяя набор правил «если — то» к набору фактов, следуя заданным процедурам.

Примечание — См. [3].

3.1.3 **выборка** (в статистике) (sample): Подмножество генеральной совокупности, состоящее из одного или нескольких выборочных единиц.

Примечания

1 В зависимости от исследуемой генеральной совокупности выборочными единицами могут быть объекты, числовые значения, а также абстрактные элементы.

2 Выборку из генеральной совокупности, подчиняющуюся нормальному распределению (2.22), гамма-распределению (2.23), экспоненциальному распределению (2.24), распределению Вейбулла (2.25), логнормальному распределению (2.26) или распределению экстремальных значений типа I (2.27) часто называют выборкой из нормального распределения, гамма-распределения, экспоненциального распределения, распределения Вейбулла, логнормального распределения или распределения экстремальных значений типа I .

[***ГОСТ Р ИСО 16269-4—2017, пункт 2.1***]

3.1.4 **знания** (knowledge): Информация об объектах, событиях, понятиях или правилах, их отношениях и свойствах, организованная для целенаправленного систематического использования.

Примечания

1 Информация может существовать в числовой или символической форме.

2 Информация — это данные, которые были контекстуализированы, чтобы их можно было интерпретировать. Данные создаются путем абстрагирования или измерений, произведенных в окружающем нас мире.

3.1.5 **пользователь** (user): Лицо или группа лиц, которые взаимодействуют с системой или получают выгоду от системы в процессе ее использования.

Примечание — См. [4], пункт 4.1.52.

3.2 Смещенность

3.2.1 **предвзятость автоматизации** (automation bias): Склонность человека отдавать предпочтение предложениям автоматизированных систем принятия решений и игнорировать противоречивую информацию, полученную без применения автоматизации, даже если она верна.

3.2.2 **смещенность** (bias): Систематическое различие в обработке определенных объектов, людей или групп по сравнению с другими.

Примечание — Обработка — это любой вид действия, включая восприятие, наблюдение, представление, предсказание или решение.

3.2.3 **когнитивная предвзятость человека** (human cognitive bias): Предвзятость (3.2.1), возникающая при обработке и интерпретации информации человеком.

Примечание — Когнитивная предвзятость человека влияет на суждения и принятие решений.

3.2.4 **предвзятость подтверждения** (confirmation bias): Тип когнитивной предвзятости человека (3.2.3), который предпочитает прогнозы систем ИИ, подтверждающие ранее существовавшие убеждения или гипотезы.

3.2.5 **удобная выборка** (convenience sample): Выборка данных, которая выбирается потому, что ее легко получить, а не потому, что она репрезентативна.

3.2.6 **смещенность данных** (data bias): Свойства данных, которые, если их не устранить, приводят к тому, что системы ИИ работают лучше или хуже для различных групп (3.2.7).

3.2.7 **группа** (group): Подмножество объектов в домене, которые связаны между собой, поскольку имеют общие характеристики.

3.2.8[1)] **статистическая смещенность** (statistical bias): Тип последовательного численного смещения в оценке относительно истинного базового значения, присущий большинству оценок.

Примечание — См. [5], пункт 3.3.9.

4 Сокращения

ИИ — искусственный интеллект;
МО — машинное обучение.

5 Обзор предвзятости (смещенности) и справедливости

5.1 Общие положения

В настоящем стандарте термин «предвзятость» определяется как систематическое различие в отношении к определенным объектам, людям или группам по сравнению с другими, в его общем значении вне контекста ИИ или МО. В социальном контексте предвзятость имеет четкую негативную коннотацию (понимание) и рассматривается как одна из основных причин дискриминации и несправедливости. Тем не менее, именно с учетом выявления систематических различий в человеческом восприятии, наблюдении и результатах, полученных в виде выходных представлений среды и ситуаций, возможно функционирование алгоритмов МО.

В настоящем стандарте термин «смещенность» используется для характеристики входных данных и составных элементов систем ИИ с точки зрения их проектирования, обучения и эксплуатации. Системы ИИ различных типов и назначения (например, для разметки, кластеризации, составления прогнозов или принятия решений) полагаются и опираются на эту смещенность для своей работы.

Для характеристики результатов работы системы ИИ или, точнее, ее возможного влияния на общество в настоящем стандарте используются термины «несправедливость» и «справедливость». Справедливость можно описать как подход, поведение или результат, основанные на уважении к установленным фактам, убеждениям и нормам и не определяемые предпочтением (фаворитизмом) или несправедливой дискриминацией.

Хотя определенная смещенность необходима для правильной работы системы ИИ, нежелательная смещенность может быть внесена в систему ИИ непреднамеренно, что приведет к несправедливым результатам работы системы.

5.2 Обзор смещенности

Системы ИИ позволяют людям по всему миру получить новый опыт и новые возможности. Системы ИИ могут использоваться для решения различных задач, таких как рекомендация книг и телевизионных передач, диагностика здоровья (прогноз наличия и тяжести медицинского состояния), подбор персонала и партнеров или обнаружение пешехода, переходящего улицу. Такие компьютерные системы для вспомогательных функций или для принятия решений имеют потенциал стать более справедливыми, также имеют риск быть менее справедливыми, чем существующие системы или чем люди, которых они дополняют или заменяют.

Системы ИИ часто обучаются на основе реальных данных; следовательно, модель МО может обучиться на основе проблемной смещенности данных или даже усилить эту смещенность. Такая смещенность может потенциально благоприятствовать или не благоприятствовать определенным группам людей, объектов, концепций или результатов. Даже если данные кажутся беспристрастными, самое тщательное кросс-функциональное обучение и тестирование все равно может привести к созданию модели МО с нежелательной смещенностью. Более того, устранение или уменьшение одного вида смещенности [например, социальной/общественной смещенности/предвзятости (societal bias)] может

[1)] Исправлена опечатка.

повлечь за собой появление или увеличение другого вида смещенности (например, статистической смещенности ***[6]***), см. положительное влияние, описанное в настоящем разделе. Смещенность может иметь негативное, позитивное или нейтральное воздействие.

Прежде чем обсуждать аспекты смещенности в системах ИИ, необходимо описать работу систем ИИ и то, что в данном контексте означает нежелательная смещенность (unwanted bias). Система ИИ может быть охарактеризована как система, использующая знания для обработки входных данных, чтобы делать прогнозы или предпринимать действия. Знания в системе ИИ часто формируются в процессе обучения на основе обучающих данных (training data); они состоят из статистических корреляций, наблюдаемых в обучающем наборе данных. Важно, чтобы непосредственно использованные рабочие (производственные) данные и данные для обучения относились к одной и той же области интересов.

Прогнозы, сделанные системами ИИ, могут быть самыми разнообразными, в зависимости от области интересов и типа системы ИИ. Однако для систем классификации полезно рассматривать прогнозы ИИ как обработку представленного ему набора входных данных и прогноз принадлежности или непринадлежности входных данных к нужному набору. В качестве простого примера можно привести прогноз, связанный с заявкой на кредит, — представляет ли заявитель приемлемый финансовый риск или нет для кредитной организации.

Оптимальная или требуемая система ИИ будет правильно предсказывать, представляет ли заявка приемлемый риск, такая система не будет способствовать систематическому исключению определенных групп. В некоторых обстоятельствах это может означать учет особенностей определенных групп, таких как этническая принадлежность и пол. Смещенность может оказывать эффект и влиять на выходной результат, который способен изменить последующие предсказания/прогнозы (вычисления). Примеры того, как определить наличие нежелательной смещенности в алгоритме в соответствии с метриками, определенными в разделе 7, приведены в приложении А.

Выявление смещенности может включать определение соответствующих критериев и анализ возможных решений, связанных с этими критериями. Учитывая конкретные критерии, в настоящем стандарте описываются методики и механизмы выявления и устранения смещенности в системах ИИ.

Классификация (тип обучения с учителем) и кластеризация (тип обучения без учителя) алгоритмов не могут функционировать без смещенности. Если все подгруппы должны рассматриваться одинаково, то эти алгоритмы должны будут маркировать все выходы одинаково (в результате чего будет получен только один класс или кластер). Однако необходимо провести исследование, чтобы оценить, является ли влияние этой смещенности положительным, нейтральным или отрицательным в соответствии с целями и задачами системы.

Примеры положительного, нейтрального и отрицательного влияния смещенности приведены ниже:

- положительный эффект. Разработчики ИИ могут внести смещенность для обеспечения справедливого результата. Например, система ИИ, используемая для найма определенного типа работников, может ввести предубеждение в отношении одного пола по сравнению с другим на этапе принятия решения, чтобы компенсировать внесенное из данных общественное предубеждение (societal bias), которое отражает их историческую недопредставленность в этой профессии;

- нейтральный эффект. Система ИИ для обработки изображений для системы самодвижущихся автомобилей может систематически ошибочно классифицировать «почтовые ящики» как «пожарные гидранты». Однако это статистическое смещение будет иметь нейтральный эффект — это будет до тех пор, пока система имеет одинаково сильное предпочтение избегать препятствия каждого типа;

- негативный эффект. Примерами негативных последствий могут быть системы найма работников с использованием ИИ, отдающие предпочтение кандидатам одного пола перед другими, и голосовые цифровые помощники, не распознающие людей с нарушениями речи. Каждый из этих случаев может иметь непреднамеренные последствия в виде ограничения возможностей тех, кого это касается. Хотя такие примеры можно отнести к категории неэтичных, смещенность — это более широкое понятие, которое применяется даже в сценариях, не имеющих негативных последствий для заинтересованных сторон, например при классификации галактик астрофизиками.

Одна из проблем, связанных с определением значимости смещенности, заключается в том, что то, что представляет собой негативное влияние, может зависеть от конкретного случая использования или области применения. Например, профилирование на основе возраста может считаться неприемлемым при принятии решений о приеме на работу. Однако возраст может играть важную роль при оценке медицинских процедур и лечения. Таким образом может быть рассмотрена возможность соответствующей адаптации к конкретному случаю использования или области применения.

В системах МО результат любой отдельной операции основывается на корреляции между признаками во входной области и ранее наблюдаемыми результатами. Любые неправильные/некорректные результаты (включая, например, автоматические решения, классификации и прогнозируемые непрерывные переменные) потенциально обусловлены плохим обобщением, результатами, использованными для обучения модели МО, и гиперпараметрами, использованными для ее калибровки. Статистическая погрешность в модели МО может быть внесена непреднамеренно или из-за погрешности в процессе сбора данных и моделирования. В символических системах ИИ когнитивные предубеждения человека могут привести к неточному определению явных знаний, например к определению правил, применимых к себе, но не к целевому пользователю, из-за предубеждений внутри группы.

Еще одна проблема, связанная со смещенностью, заключается в легкости ее распространения в системе, после чего ее может быть сложно распознать и устранить. Примером может служить ситуация, когда данные отражают предубеждение, существующее в обществе, и это предубеждение становится частью новой системы ИИ, которая затем распространяет первоначальное предубеждение/предвзятость.

Организации могут учитывать риск нежелательной смещенности в наборах данных и алгоритмах, включая те, которые на первый взгляд кажутся безобидными и безопасными. Кроме того, после того как были предприняты попытки устранить нежелательную смещенность, непреднамеренная категоризация и неумелые алгоритмы могут закрепить навсегда или усилить существующую смещенность. Как следствие, устранение нежелательной смещенности не является процессом «настроил и забыл».

Например, алгоритм рассмотрения резюме, отдающий предпочтение кандидатам с многолетним непрерывным стажем работы, автоматически поставит в невыгодное положение тех, кто возвращается на работу после перерыва по уходу за ребенком. Аналогичный алгоритм может также понизить рейтинг случайных работников, чья трудовая история состоит из множества коротких контрактов у самых разных работодателей: эта характеристика может быть неверно истолкована как негативная. Тщательная переоценка вновь достигнутых результатов может последовать за любым нежелательным снижением смещенности и переобучением алгоритма.

Чем более автоматизирована система и чем менее эффективен человеческий надзор, тем выше вероятность непреднамеренных негативных последствий. Ситуация усугубляется, когда несколько приложений ИИ способствуют автоматизации той или иной задачи. В таких многоприкладных системах ИИ организации, внедряющие их, могут ожидать повышенных требований и спроса на прозрачность и объяснимость результатов.

5.3 Обзор вопросов справедливости

Справедливость — это понятие, которое отличается от предвзятости (смещенности), но связано с ней. Справедливость может характеризоваться воздействием системы ИИ на отдельных людей, группы людей, организации и общества, на которые эта система оказывает влияние. Однако невозможно гарантировать всеобщую справедливость. Справедливость — сложное понятие, очень контекстуальное и иногда оспариваемое, варьирующееся между культурами, поколениями, географическими регионами и политическими взглядами. То, что считается справедливым, может быть непоследовательным в этих контекстах. Таким образом, в настоящем стандарте не дается определение термина «справедливость», поскольку он в значительной степени зависит от социальных и этических условий.

Даже в контексте ИИ трудно дать определение справедливости таким образом, чтобы оно было одинаково применимо ко всем системам ИИ во всех контекстах. Система ИИ может потенциально влиять на отдельных людей, группы людей, организации и общества многими нежелательными способами. Общие категории негативных воздействий, которые могут восприниматься как «несправедливые», включают:

- несправедливое предоставление или выдача: происходит, когда система ИИ несправедливо предоставляет или не выдает возможности или ресурсы таким образом, что это негативно сказывается на одних сторонах по сравнению с другими;
- несправедливое качество услуг: возникает, когда система ИИ работает менее эффективно для одних сторон, чем для других, даже если нет возможностей или ресурсов для предоставления или удержания;
- стереотипизация: происходит, когда система ИИ усиливает существующие общественные стереотипы;

- унижение или уничижение: происходит, когда система ИИ ведет себя унизительным или уничижительным образом;

- «чрезмерное» или «недостаточное» представление и игнорирование (удаление из процесса вычислений, оставление без внимания): происходит, когда система ИИ чрезмерно или недостаточно представляет одни стороны в сравнении с другими, или даже не отражает их существование.

Смещенность — это лишь один из многих элементов, которые могут влиять на справедливость. Было отмечено, что смещенные входные данные не всегда приводят к несправедливым прогнозам и действиям, а несправедливые прогнозы и действия не всегда вызваны смещенностью.

Примером предвзятой системы принятия решений, которая, тем не менее, может считаться справедливой, является политика приема на работу в университет, которая предвзята в пользу людей с соответствующей квалификацией, поскольку она пропускает и нанимает гораздо большую долю обладателей соответствующей квалификации, чем доля обладателей соответствующей квалификации среди общего населения. До тех пор, пока определение соответствующей квалификации не дискриминирует определенные демографические группы, такая система может считаться справедливой.

Примером непредвзятой системы, которую можно считать несправедливой, является политика, которая без разбора отвергает всех кандидатов. Такая политика действительно была бы непредвзятой, поскольку не делала бы различий между какими-либо категориями. Но она будет восприниматься как несправедливая людьми с соответствующей квалификацией.

В настоящем стандарте проводится различие между предвзятостью (смещенностью) и справедливостью. Предвзятость (смещенность) может быть общественной/социальной (societal) или статистической (statistical), может отражаться в различных компонентах системы или возникать из них (см. раздел 6) и может вноситься или распространяться на различных этапах жизненного цикла разработки и внедрения ИИ (см. раздел 8).

Достижение справедливости в системах ИИ часто означает принятие компромиссных решений. В некоторых случаях различные заинтересованные стороны могут иметь законно противоречащие друг другу приоритеты, которые невозможно примирить с помощью альтернативного проектирования системы. В качестве примера рассмотрим систему ИИ, которая принимает решение о присуждении стипендий некоторым абитуриентам выпускных программ в университете. Заинтересованное лицо в приемной комиссии, занимающееся вопросами разнообразия, хочет, чтобы система искусственного интеллекта обеспечивала справедливое распределение таких стипендий среди абитуриентов из различных географических регионов. Профессор, который является другой заинтересованной стороной, хочет, чтобы стипендию получил конкретный достойный студент, заинтересованный в определенной области исследований. В таком случае существует вероятность того, что система ИИ откажет достойному кандидату из определенного региона, чтобы удовлетворить цели исследования. Таким образом, удовлетворить ожидания справедливости всех заинтересованных сторон не всегда возможно. Поэтому важно быть понятным и прозрачным в отношении этих приоритетов и любых лежащих в их основе предположений, чтобы правильно выбрать соответствующие метрики (см. раздел 7).

6 Источники нежелательной смещенности в системах ИИ

6.1 Общие положения

В данном пункте описываются возможные источники нежелательной смещенности в системах ИИ. К ним относятся когнитивная предвзятость человека, предвзятость данных и предвзятость, вызванная инженерными решениями. На рисунке 1 показана взаимосвязь между этими высокоуровневыми группами смещенности. Когнитивные предвзятости человека (см. 6.2) могут вызвать смещенность, вносимую инженерными решениями (см. 6.4) или смещенностью данных (см. 6.3).

Например, письменный или устный язык содержит общественные предубеждения, которые могут быть усилены моделями встраивания слов [7]. Поскольку общественная предвзятость отражается в существующем языке, который используется в качестве обучающих данных, она в свою очередь вызывает нерепрезентативную смещенность данных выборки (см. 6.3.4), что может привести к нежелательной смещенности. Эта взаимосвязь показана на рисунке 2.

Рисунок 1 — Взаимосвязь между высокоуровневыми группами смещенности

Рисунок 2 — Пример общественного предубеждения, проявляющегося в виде нежелательной смещенности

В системах, скорее всего, будет одновременно присутствовать несколько источников смещенности. Анализ системы с целью выявления одного источника смещенности вряд ли позволит выявить все. В том же примере для обработки естественного языка используется несколько моделей. Результаты модели встраивания слов, на которые может повлиять нерепрезентативность выборки, затем обрабатываются вторичной моделью. В этом случае вторичная модель уязвима к смещенности при разработке параметров, поскольку был сделан выбор в пользу использования встраивания слов в качестве особенности этой модели.

Не все источники смещенности начинаются с когнитивных предубеждений человека, смещенность может быть вызвана исключительно характеристиками данных. Например, датчики, подключенные к системе, могут выйти из строя и выдать сигналы, которые можно считать отклонениями (см. 6.3.10). Эти данные, когда они используются для обучения или обучения с подкреплением, могут внести нежелательную погрешность. Это показано на рисунке 3.

Рисунок 3 — Пример характеристики данных, проявляющихся в виде нежелательной смещенности

6.2 Когнитивная предвзятость человека

6.2.1 Общие положения

Когнитивная предвзятость человека (human cognitive biases) может проявляться по-разному, как осознанно, так и бессознательно, на нее влияют данные, информация и опыт, доступные для принятия решений [8]. Мышление часто основано на непознанных, непрозрачных процессах, которые заставляют человека принимать решения, не всегда понимая, чем они обусловлены. Эти когнитивные предубеждения человека влияют на решения о сборе и обработке данных, проектировании системы, обучении моделей и другие решения о разработке, которые принимают люди, а также на решения о том, как используется система.

6.2.2 Предвзятость при автоматизации

ИИ способствует автоматизации анализа и принятия решений в различных системах, например, в беспилотных автомобилях и системах здравоохранения, что может вызвать предвзятости при автоматизации (automation bias). Предвзятость при автоматизации возникает, когда человек, принимающий решение, отдает предпочтение рекомендациям, сделанным автоматизированной системой принятия решений, по сравнению с информацией, полученной без автоматизации, даже если автоматизация допускает ошибки.

6.2.3 Смещенность групповой атрибуции

Смещенность групповой атрибуции (group attribution bias) возникает, когда человек предполагает, что то, что верно для отдельного человека или объекта, также верно для всех или всех объектов в этой группе. Например, влияние групповой атрибуции может усугубиться, если для сбора данных используется удобная выборка. В нерепрезентативной выборке могут быть сделаны атрибуции, которые не отражают реальность. Это также является одним из видов статистической смещенности.

6.2.4 Неявная смещенность

Неявная смещенность (implicit bias) возникает, когда человек делает ассоциации или предположения, основанные на его ментальных моделях и воспоминаниях. Например, при создании классификатора для идентификации свадебных фотографий инженер может использовать в качестве признака наличие белого платья на фотографии. Однако белые платья были в обычае только в определенные эпохи и в определенных культурах.

6.2.5 Предвзятость подтверждения

Предвзятость подтверждения (confirmation bias) возникает, когда гипотезы, независимо от их истинности, с большей вероятностью подтверждаются в результате преднамеренной или непреднамеренной интерпретации информации.

Например, разработчики в области МО могут непреднамеренно собирать или размечать данные таким образом, чтобы повлиять на результат, подтверждающий их существующие убеждения. Предвзятость подтверждения — это одна из форм неявной смещенности.

Предвзятость экспериментатора (experimenter's bias) — это форма предвзятости подтверждения, когда экспериментатор продолжает обучение моделей до тех пор, пока не подтвердится ранее существовавшая гипотеза.

Когнитивная предвзятость человека, в частности предвзятость подтверждения, могут вызывать различные другие смещенности, например смещенность отбора (см. 6.3.2) или смещенность в разметке данных (см. 6.3.3).

Другим примером является предвзятость «Что видишь, то и есть» (WYSIATI). Это происходит, когда человек ищет информацию, подтверждающую его убеждения, игнорирует противоречивую информацию и делает выводы на основе того, что ему знакомо [9].

6.2.6 Внутригрупповая предвзятость

Внутригрупповая предвзятость (in-group bias) возникает при проявлении предпочтения (пристрастности) к собственной группе или собственным характеристикам. Например, если персонал, осуществляющий тестирование или оценку, состоит из друзей, родственников или коллег разработчика системы, то внутригрупповая предвзятость может сделать недействительным тестирование продукта или набора данных. Это может выражаться в оценке других людей, распределении ресурсов и многими другими способами.

Было установлено, что люди стремятся сделать больше внутренних (диспозиционных) атрибуций для событий, которые позитивно отражаются на группах, к которым они принадлежат, и больше внешних (ситуационных) атрибуций для событий, которые негативно отражаются на их группах.

6.2.7 Предвзятость однородности внешней группы

Предвзятость в отношении однородности внешней группы (out-group homogeneity bias) возникает, когда при сравнении установок, ценностей, черт личности и других характеристик члены внешней группы воспринимаются как более похожие, чем члены внутренней группы. Например, европейцы могут восприниматься американцами как одна однородная, единообразная группа, и наоборот. Однако в каждой группе можно выделить множество подгрупп и специфических черт, что доказывает большое разнообразие, существующее в реальности.

Эффект однородности внешней группы (out-group homogeneity effect) — это восприятие индивидом членов внешней группы как более похожих друг на друга, чем члены внутренней группы, например, «они похожи, а мы разные». Термин «эффект однородности внешней группы» или «относительная однородность внешней группы» явно противопоставляется термину «однородность внешней группы» в целом, последний означает воспринимаемую изменчивость внешней группы, не связанную с восприятием внутренней группы.

Эффект однородности внешней группы является частью более широкой области исследований, изучающих воспринимаемую групповую изменчивость. Эта область включает эффекты внутригрупповой однородности, а также эффекты внегрупповой однородности. Эффект внутригрупповой однородности возникает, когда члены группы воспринимаются как схожие в отношении положительных характеристик. В этой области исследований также рассматриваются эффекты воспринимаемой групповой изменчивости, которые не связаны с членством в группе или вне группы, например эффекты, связанные с властью, статусом и размером групп.

Эффект однородности внешней группы был обнаружен на примере широкого спектра различных социальных групп, от политических и этнических групп до возрастных и гендерных групп.

6.2.8 Общественная или социальная предвзятость

Общественная предвзятость (societal bias) возникает, когда многие люди в обществе имеют схожие когнитивные предубеждения (сознательные или бессознательные). Следовательно, эта предвзятость может быть закодирована, воспроизведена и закреплена в политике организаций.

Она проявляется в МО, когда модели учатся или усиливают существующие, исторические модели предвзятости в наборах данных. Данная общественная предвзятость исходит от общества в целом и может быть тесно связана с когнитивной предвзятостью или статистической смещенностью. Она проявляется в виде доступных данных об обществе, которые отражают исторические закономерности. Общественную предвзятость также можно считать одним из видов смещенности данных (см. 6.3).

Общественная предвзятость также проявляется, когда культурные предположения о данных применяются без учета межкультурных различий. Например, многие группы относятся к данным по геномике как к полностью светским, но некоторые группы считают, что геномика также содержит священные или духовные свойства. Модель, построенная на этих данных, может сбалансированно предсказывать заболевания в разных популяциях. Однако если эти данные затрагивают социальные группы, которые считают эти данные священными, а разработчик не признает или не учитывает эти культурные различия, модель может укоренить общественное предубеждение, независимо от численных результатов.

Одним из примеров общественного предубеждения является ситуация, когда исторические данные не соответствуют сделанным выводам, возможно, укрепляя общепринятые, но неточные социальные взгляды. Например, прогнозирование того, совершит ли заключенный новое преступление после условно-досрочного освобождения (т.е. уровень рецидива), зависит от наличия данных о том, какие предыдущие заключенные совершили какие виды преступлений [10], если таковые были после того, как они также были условно-досрочно освобождены. Однако имеющиеся данные ограничиваются бывшими заключенными, которые были арестованы или осуждены за преступление после освобождения. Документально подтверждено, что полицейские аресты и судебные приговоры в значительной степени зависят от отношения к этнической принадлежности, бедности и предыдущим арестам. Например, любой систематический арест и осуждение определенной группы людей приведет к систематической чрезмерной классификации рецидива среди этой группы заключенных.

Системная предвзятость (systemic bias), также называемая институциональной предвзятостью, — это форма общественной предвзятости, встречающаяся в системах. Системная предвзятость — это присущая социотехнической системе или процессу тенденция поддерживать определенные результаты.

Термин системная предвзятость исторически используется в контексте человеческих систем и процессов, действующих в организациях или в обществе или культуре, и широко обсуждается в области экономики производственной организации.

Например, системная предвзятость играет роль в системном расизме. Системный расизм — это форма расизма, которая может быть встроена в общество, определенную культуру или организацию.

6.2.9 Проектирование систем на основе правил

Опыт разработчиков и советы экспертов могут оказать значительное влияние на проектирование системы на основе правил (rule-based system design), а также потенциально привнести различные формы когнитивной предвзятости человека. Разработчик может, например, ввести прямое правило, основанное на предположении о прибыли, которое делает разделение в популяции таким образом, что применяются отдельные модели для людей, получающих регулярный доход на свои банковские счета, и для тех, которые его не получают. Такое разделение может привести к предвзятому отношению к тем, кто работает на себя, по сравнению с теми, кто работает на третьих лиц. Правило также может несправедливо дискриминировать различные демографические группы населения, если существует связь между типом занятости и социально-демографическими характеристиками в конкретном географическом месте.

6.2.10 Предвзятость требований

Формирование требований создает условия для проявлений когнитивной предвзятости человека (см. 6.2). Например, неявные предположения о возможностях аппаратных средств, сделанные разработчиками ИИ с высоким социально-экономическим статусом, не обязательно будут справедливы для всех пользователей системы ИИ. В целом, когнитивные предвзятости человека будут приводить к тому, что разработчики ИИ будут брать за основу и использовать условия, схожие с их собственными, которые не являются репрезентативными для общей целевой группы пользователей. Примеры стратегий снижения предвзятости при разработке требований приведены в разделе 8.

Количественные характеристики, оптимизируемые в процессе обучения модели, также могут привнести в систему предвзятость требований (requirements bias). Упрощенное сведение требований к уравнению полезности может создать предвзятость (смещение) требований.

6.3 Смещенность данных

6.3.1 Общие положения

Одним из основных источников смещенности являются данные, используемые для обучения и разработки систем ИИ. В 6.3 подробно описаны конкретные ситуации и сценарии, при которых данные могут быть смещенными. Смещенность данных (data bias) возникает в результате технических решений и ограничений, а также может быть вызвана когнитивной предвзятостью человека, выбранной методикой обучения или вариациями в инфраструктуре обучения. Эти источники не являются исключительными для систем ИИ и могут быть обнаружены в других приложениях или системах. Однако то, как они проявляются в системах ИИ, соответствует определенным закономерностям. Например, смещенность, вызванная обучающим набором данных, может объясняться неправильным применением или игнорированием статистических методов и правил.

6.3.2 Статистическая смещенность

6.3.2.1 Смещенность отбора

6.3.2.1.1 Общие положения

Смещенность отбора (selection bias) возникает, когда выборка подобрана таким образом, что она не отражает реальное положение вещей и распространение элементов выборки в реальном мире. Смещенность отбора может быть связана с когнитивной предвзятостью человека в процессе отбора данных (см. 6.2).

6.3.2.1.2 Смещенность выборки

Смещенность выборки (sampling bias) возникает, когда записи данных не отбираются случайным образом из предполагаемой совокупности.

Если набор данных смещен (необъективен) по количеству выборок, взятых из различных групп, то модель не будет точно отражать среду, в которой она будет эксплуатироваться. Например, система распознавания лиц, обученная на людях только одного пола или только одной расы, скорее всего, не сможет так же успешно распознавать лица людей, которых нет в наборе данных для обучения ***[11]***.

6.3.2.1.3 Смещенность по охвату

Смещенность по охвату (coverage bias) или смещенность по покрытию происходит, когда популяция, представленная в наборе данных, не совпадает с популяцией, в отношении которой модель МО

делает прогнозы. Например, если построить модель МО для предсказания удовольствия от просмотра драматических фильмов на основе опроса зрителей комедийных фильмов, она явно будет иметь смещенность по охвату, которая может быть весьма существенной.

6.3.2.1.4 Смещенность, связанная с отсутствием ответа

Смещенность, связанная с отсутствием ответа (non-response bias), также называемая смещенностью участия (participation bias), возникает, когда люди из определенных групп отказываются от участия в опросах с разной частотой, чем респонденты из других групп.

6.3.2.2 Спутывающие переменные

Спутывающая переменная (confounding variable) — это переменная, которая влияет как на зависимую, так и на независимую переменную, вызывая ложную связь. Из-за этого предполагаемая связь между двумя переменными может быть доказана как частично или полностью ложная.

6.3.2.3 Набор данных, который не имеет нормального распределения

Большинство статистических методов предполагают, что набор данных имеет нормальное распределение. Однако если набор данных имеет другое распределение (например, хи-квадрат, бета, Лоренца, Коши, Вейбулла или Парето), результаты могут быть необъективными и вводящими в заблуждение (non-normality).

6.3.3 Разметка данных и процесс разметки

Сам процесс разметки (labelling process) потенциально привносит в данные когнитивные или общественные предвзятости/смещения (см. 6.2). ***Например, принимая решение о классификации людей на стариков и молодых, имеющих официальный заработок или нет, здоровых и инвалидов, люди распределяются по дискретным категориям, которые не обязательно отражают всю реальность, которая моделируется[1)].*** Могут быть выбраны ярлыки (метки), которые могут быть слишком широко интерпретированы или которые сводят непрерывный спектр к бинарной переменной. В других случаях процесс разметки естественным образом попадает в дискретное пространство, но истинные метки недоступны. В таких случаях часто используются приближенные значения, которые коррелируют с истинными метками и считаются достаточно близкими для большинства целей. Если неточности, вносимые таким прокси, не являются случайными, они могут внести в систему погрешность. Например, система искусственного интеллекта, рекомендующая право на условно-досрочное освобождение (см. 6.2.8), может быть также описана как в целом не имеющая доступа к информации о том, могли ли люди, не вышедшие на свободу, совершить новые преступления.

Наконец, сам процесс разметки может быть изначально несовершенным. В процессе разметки данных в них могут быть внесены когнитивные предвзятости людей, занимающихся разметкой данных. Также возможно, что такая предвзятость будет включена в инструкции по разметке.

6.3.4 Нерепрезентативная выборка

Смещенность может проявляться несколькими способами во время подготовки обучающих данных, как результат когнитивной предвзятости человека (см. 6.2), или из-за смещенности выборки или охвата (см. 6.3.2). Иногда все доступные наборы данных имеют свойства, унаследованные от когнитивной предвзятости человека, который их создал. Когнитивная предвзятость человека в процессе отбора может препятствовать использованию или созданию объективных (несмещенных) наборов данных. Нерепрезентативная выборка (non-representative sampling) является примером смещенной подготовки обучающих данных. Большинство методов моделирования рассматривают обучающие данные как истинную и точную картину моделируемого явления. Если набор данных не является репрезентативным для предполагаемой среды эксплуатации (домена), то модель может быть наделена смещенностью, основанной на нерепрезентативности набора данных.

Репрезентативность может принимать различные формы в разных областях применения. Например, в области распознавания лиц существует несколько различных вариантов того, как набор данных может стать нерепрезентативным в отношении такого атрибута, как цвет кожи. Количество изображений людей с определенным цветом кожи, условия освещения изображений и относительная энтропия изображений людей с одним цветом кожи являются примерами того, как нерепрезентативный набор данных может привнести смещенность в модель.

[1)] Перечень примеров скорректирован для учета особенностей, характерных для Российской Федерации.

В данных могут присутствовать признаки, которые могут позволить модели МО косвенно вывести принадлежность к группе, даже если сами признаки принадлежности к группе не входят во входные данные модели МО (см. 8.3.3.1).

6.3.5 Недостающие характеристики и метки

Данные реального мира редко бывают полными. В частности, полные признаки и особенности часто отсутствуют в отдельных обучающих выборках. Если частота отсутствующих признаков выше для одной группы, чем для другой, то это представляет собой еще один источник смещенности. Например, история болезни для определенных групп людей часто менее полна по сравнению с другими группами из-за более фрагментарного лечения, которое они получают в среднем. Такой дисбаланс в качестве данных может привести к ухудшению качества медицинских прогнозов.

6.3.6 Обработка данных

Смещенность также может возникнуть в результате предварительной (или последующей) обработки данных, даже если исходные данные не привели бы к какой-либо смещенности. Например, выравнивание недостающих значений, исправление ошибок, удаление выпадающих значений или принятие определенных моделей распределения данных также может привести к смещенности в работе системы ИИ. Это может быть вызвано когнитивной предвзятостью человека (см. 6.2).

6.3.7 Парадокс Симпсона

Парадокс Симпсона (Simpson's paradox) проявляется в том случае, когда тенденция, наблюдаемая в отдельных группах данных, меняется на противоположную, когда группы данных объединяются. Причиной такого проявления обычно является различное соотношение весов отдельных групп.

6.3.8 Агрегирование данных

Агрегирование данных (aggregating data), охватывающих различные группы объектов, которые имеют различные статистические распределения, может внести смещенность в данные, используемые для обучения систем ИИ ***[12]***. Это может быть вызвано когнитивной предвзятостью человека, такой как предвзятость однородности внешней группы.

6.3.9 Распределенное обучение

Из-за соображений конфиденциальности и соответствующих нормативных требований обучение, которое приближено к источнику данных, может получить широкое распространение с использованием распределенных методологий и приемов. Распределенное МО (distributed training) может привнести свою собственную причину смещенности данных, поскольку различные источники данных могут иметь различное распределение признаков. Если все источники данных, которые в совокупности влияют на полноту признаков, не участвуют в обучении, может возникнуть смещенность, соответствующая пространству признаков, не участвующих в обучении. Неучастие может произойти из-за проблем с сетью, низкой способности вычислительных устройств для соответствующих источников данных или отсутствия выбора источника данных.

6.3.10 Другие источники смещенности данных

Данные и любая разметка данных могут быть искажены артефактами или другими вредными воздействиями. Такая смещенность будет рассматриваться алгоритмом ИИ как часть обобщаемой модели и это приведет к нежелательным результатам. Например:

- выбросы — это экстремальные значения данных, которые, если они реальны, представляют собой события с очень низкой вероятностью в моделируемых данных;

- шум — это искажение, которое характеризуется статистически распределенной вариацией физической величины. Шум вызван стохастическими процессами (stochastic processes) и не может быть описан детерминировано. Шум может оказывать негативное влияние на модель, если происходит чрезмерная подгонка (настройка). Кроме того, искусственно созданный шум может быть использован для создания неблагоприятных примеров, которые приведут к нежелательным результатам.

6.4 Смещенность, вызванная инженерными решениями

6.4.1 Общие положения

Архитектуры моделей МО, включающие все спецификации моделей, параметры и разработанные специалистами (вручную) функции, могут быть смещенными по целому ряду причин. Этому могут способствовать смещенность данных и когнитивная предвзятость человека.

6.4.2 Разработка (инжиниринг) признаков

В процессе разработки признаков (feature engineering) при построении модели МО разработчики ИИ могут напрямую использовать любые входные элементы или создавать сложные компоненты для модели МО из входных элементов таким образом, что они могут быть линейными или нелинейными комбинациями некоторых входных элементов. Такие этапы, как кодирование, преобразование типа данных, снижение размерности и выбор признаков зависят от выбора разработчика ИИ и могут внести смещенность в модель МО.

Например, разработчик ИИ может выбрать для представления человеческого роста категориальные значения, такие как высокий, средний или низкий, а затем выбрать диапазоны таким образом, что большинство представителей одного пола попадут в категорию среднего и низкого роста, а большинство представителей другого пола — в категорию высокого и среднего роста. Это может внести в модель нежелательную смещенность. Другой пример — разработчик ИИ может использовать сложную характеристику индекса массы тела, составленную из роста и веса человека, а затем создать модель, чтобы использовать характеристику индекса массы тела, а не исходные характеристики роста и веса. Это может внести смещенность, несправедливую по отношению к некоторым группам, таким как профессиональные борцы сумо и тяжелоатлеты.

Иногда скрытые или неявные корреляции между признаками (элементами) могут стать заметными из-за недостаточной настройки или недостаточного количества признаков модели. Это может проявиться в виде нежелательной смещенности в прогнозах системы.

6.4.3 Выбор алгоритма

Выбор алгоритмов МО (algorithm selection), встроенных в систему ИИ, может внести нежелательную смещенность в прогнозы, сделанные системой. Это происходит потому, что тип используемого алгоритма вносит изменения в эффективность и производительность (performance) модели МО.

В простейшем примере это может быть использование линейной модели для решения нелинейной задачи. В более сложном примере существуют различные возможные конфигурации моделей долговременной и кратковременной памяти, — такие модели могут состоять из нескольких слоев. Это напрямую влияет на сложность функции, которую сеть способна аппроксимировать. Другие архитектуры нейронных сетей, такие как модели трансформатор-энкодер-декодер (transformer-encoder-decoder models), имеют функциональность, которая может внести нежелательную смещенность в прогнозы, сделанные системой.

Внутри модели МО может быть много подмоделей, которые могут взаимодействовать с линейной комбинацией или более сложной комбинацией подмоделей. Таким образом, может возникнуть множество сложных проблем, включая нежелательную смещенность в прогнозах системы ИИ.

Например, в модели МО для системы ответов на вопросы на естественном языке может быть комбинация модели предсказания предикатов, модели идентификации значений, модели связывания предикатов и значений и модели идентификации ограничений. Способ комбинирования или последовательной работы этих субмоделей может внести нежелательную смещенность в прогнозы системы.

Градиентный бустинг (gradient-boosting) также может быть использован для объединения набора подмоделей МО в единую сильную обучающую систему методом итерации. Однако ансамбль (совокупность) таких подмоделей может внести нежелательную смещенность в конечные прогнозы. Например, модель МО может использовать последовательное построение неглубоких деревьев регрессии (shallow regression trees) для формирования ансамбля и выдавать прогноз как сумму вероятностей предсказаний деревьев. Способ построения ансамбля может внести смещенность в систему.

6.4.4 Настройка гиперпараметров

При создании модели МО выбор способа проектирования определяет архитектуру модели. Часто оптимальная архитектура модели развивается путем настройки гиперпараметров. Гиперпараметры включают количество слоев сети, количество нейронов в слое (также называемое шириной каждого слоя), скорость обучения для градиентного спуска (gradient descent), степень полиномов (polynomials), используемых для линейной модели, количество деревьев в случайном лесу и т. д.

Гиперпараметры определяют структуру модели и не могут быть непосредственно обучены по данным, как параметры модели. Таким образом, гиперпараметры влияют на функционирование и точность модели и, следовательно, потенциально могут привести к смещению (предвзятости).

Существует множество возможных функций активации (activation functions) для нейронной сети. Выбор функций активации может повлиять на точность и прогнозы, сделанные моделью МО. Это может проявиться в виде предвзятости в прогнозах, сделанных системой.

Кроме того, часто необходимо выбрать порог принятия решения (pick a decision threshold) для выполнения того или иного действия данной моделью. Зачастую такие пороги устанавливаются вручную. Таким образом, если модель обновляется на основе новых данных, ранее установленный вручную порог может стать некорректным или привести к предвзятости в прогнозах. Это особенно важно для динамических систем.

6.4.5 Информативность

Для некоторых групп карта взаимосвязей (mapping) между входными элементами, присутствующими в данных, и выходными элементами является более сложной для изучения. Это происходит, когда некоторые признаки высокоинформативны для одной группы, а другой набор признаков высокоинформативен для другой группы. В таком случае модель, в которой доступен только один набор признаков, может быть смещенной по отношению к группе, взаимосвязи которой трудно изучить на основе имеющихся данных. Эта концепция применима как при обучении, так и при оценке модели. Выраженность или экспрессивность модели (model expressiveness) (см. 6.4.7.2) также является фактором информативности.

6.4.6 Смещенность модели

Учитывая, что в МО часто используются такие возможности, как метод максимального правдоподобия (maximum likelihood estimator) для определения параметров, то если в данных присутствует перекос или недостаточная репрезентативность, метод максимального правдоподобия имеет тенденцию усиливать любую смещенность (предвзятость), лежащую в основе распределения. Например, если распределение мужчин и женщин в наборе данных составляет 60 % мужчин и 40 % женщин, модель может представить этот перекос в виде 80 % мужчин и 20 % женщин, используя пороговые значения, которые не учитывают исходную смещенность. Нисходящие функции активации, такие как сигмоидная функция (sigmoid function), могут усиливать небольшие различия в характеристиках, которые являются результатом смещенности данных.

6.4.7 Взаимодействие моделей

6.4.7.1 Общие положения

Структура модели может создавать необъективные прогнозы. Например, предположим, что переменные X и Y имеют значение для прогнозирования результатов в двух группах, но являются независимыми в одной группе и взаимодействующими в другой. Модель, в которой эти две переменные присутствуют, но не могут быть изолированы, потенциально даст смещенные результаты.

6.4.7.2 Выраженность (экспрессивность) модели

Модели обладают различной способностью к проявлению (экспрессивностью) (expressive capacity), и некоторые из них охватывают более широкое разнообразие функций, чем другие. Количество и характер параметров в модели, а также топология нейронной сети могут влиять на экспрессивность модели. Любая особенность, которая по-разному влияет на экспрессивность модели в разных группах, потенциально может вызвать смещенность.

Архитектуры моделей, допускающие рекурсию, также могут обеспечить большую выраженность (экспрессивность). Свойства некоторых групп могут быть полностью поняты через статическое представление текущего состояния. Те же свойства других групп могут быть поняты как результат последовательности состояний. В этом случае нерекуррентная модель будет лучше работать для первой группы, чем для второй.

7 Оценка предвзятости (смещенности) и справедливости в системах искусственного интеллекта

7.1 Общие положения

При разработке и внедрении системы ИИ важно знать о возможной предвзятости (включая статистическую смещенность и общественную предвзятость), которая может привести к несправедливому поведению системы. Одним из способов обнаружения признаков нежелательной предвзятости

является оценка результатов работы системы с помощью одной или нескольких метрик справедливости (fairness metrics). Нежелательная предвзятость, обнаруженная с помощью такой оценки, может быть устранена с помощью методов, описанных в разделе 8.

Метрики статистической смещенности (metrics of statistical bias) направлены на оценку различий между средними наблюдаемыми значениями и истинными значениями. С распространением систем ИИ и озабоченностью по поводу их справедливости растет понимание того, что такие метрики статистической смещенности недостаточны для выявления несправедливого или дискриминационного поведения. Это привело к разработке метрик ***[13]***, цель которых отразить различные представления о справедливости.

Такие метрики описаны в литературе по «алгоритмической справедливости» («algorithmic fairness») ***[14]*** и называются «метриками справедливости» («fairness metrics») или «метриками алгоритмической справедливости» («metrics of algorithmic fairness»). Например, некоторые метрики справедливости предназначены для сравнения различных типов ошибок между различными группами людей.

Следует отметить, что не существует однозначного соответствия между широким понятием предвзятости (как определено в настоящем стандарте) и статистическими метриками предвзятости. Также нет однозначного соответствия между широким понятием справедливости (как описано в 5.3) и метриками справедливости. Основной проблемой остается определение метрик, наиболее подходящих в любом конкретном контексте ***[15]***.

На сегодняшний день большинство работ по метрикам справедливости сосредоточено на справедливости систем ИИ, основанных на классификации или регрессии (classification- or regression-based AI systems), по отношению к группам, определенным по одному или нескольким демографическим признакам. В этом пункте представлены подходы к оценке предвзятости и справедливости систем ИИ, основанных на классификации. Аналогичные концепции существуют и для регрессионных систем ИИ (примеры см. в ***[16]***, ***[17]***).

Предвзятость в системах классификации может быть обнаружена путем измерения различных типов ошибок по отношению к различным группам. Подход, при котором данные разделяются на обучающие, проверочные и тестовые наборы, дополняется делением на подразделы (subdividing) каждого из этих наборов данных на основе характеристик, в отношении которых ожидается, что система будет справедлива. Если существует несколько характеристик, имеющих отношение к обнаружению возможных смещений (предвзятости) в конкретной системе, то эти характеристики могут рассматриваться как независимые или как пересекающиеся. Например, система, которая является объективной по отношению к полу и расе независимо друг от друга, может быть предвзятой по отношению к определенной комбинации этих двух характеристик.

Перед тестированием можно четко сформулировать цели обеспечения справедливости, что включает определение соответствующих демографических характеристик, выбор и обоснование метрик справедливости, которые будут использоваться для выявления предвзятости, и установление допустимой границы различий («дельты»).

После того, как данные были соответствующим образом разделены, предварительно определенные метрики справедливости рассчитываются для каждой группы и проводится сравнение между группами. Система классификации может считаться достаточно справедливой (или «несмещенной», «непредвзятой») в отношении соответствующих характеристик, если основанные на метрике измерения между группами находятся в пределах достаточно малой дельты.

7.2 Матрица ошибок

Матрица ошибок ***[18]*** (см. рисунок 4) — это инструмент, который может быть использован для оценки эффективности классификатора. Он сообщает о количестве ложноотрицательных, ложноположительных, истинно положительных и истинно отрицательных результатов и включает дополнительные критерии эффективности, полученные на основе этих значений. Матрица ошибок содержит и сравнивает несколько метрик, она позволяет детально проанализировать эффективность классификатора и обойти или выявить недостатки отдельных метрик.

		Истинные условия			
		Условие положительно	Условие отрицательно	Распространенность (Prevalence) $= \frac{\sum \text{условие положительно}}{\sum \text{общая популяция}}$	Точность (Accuracy, *ACC*) $= \frac{\sum TP + \sum TN}{\sum \text{общая популяция}}$
Предсказанные условия	Предсказание положительно	Истинно положительные (true positive, *TP*) Сила (Power)	Ложноположительные (false positive, *FP*) Ошибка 1-го типа	Положительно предсказательные значения, Точность, Актуальность (Positive Predictive Value, *PPV*, Precission, Relevance) $= \frac{\sum TP}{\sum \text{предсказание положительно}}$	Коэффициент ложного обнаружения (False Discovery Rate, *FDR*) $= \frac{\sum FP}{\sum \text{предсказание положительно}}$
	Предсказание отрицательно	Ложноотрицательные (false negative, *FN*) Ошибка 2-го типа	Истинно отрицательные (true negative, *TN*)	Коэффициент ложных пропусков (False Omission Rate, *FOR*) $= \frac{\sum FN}{\sum \text{предсказание отрицательно}}$	Отрицательно предсказательные значения (Negative Prediction Value, *NPV*, Separation Ability) $= \frac{\sum TN}{\sum \text{предсказание отрицательно}}$
Общая популяция		Истинно положительная пропорция, Полнота, Чувствительность (True Positive Rate, *TPR*, Recall, Sensitivity) $= \frac{\sum TP}{\sum \text{условие положительно}}$	Ложноположительная пропорция, Вероятность ложного срабатывания (False Positive Rate, *FPR*, Fall-out, Probability False Alarm) $= \frac{\sum FP}{\sum \text{условие отрицательно}}$	Положительное соотношение правдоподобия (Positive Likelihood Ratio, *LR*+) $= \frac{TPR}{FPR}$	Коэффициент диагностической вероятности (Diagnostic Odds Rate, *DOR*) $= \frac{LR+}{LR-}$
		Ложноотрицательная пропорция, Процент пропусков (False Negative Rate, сокр. *FNR*, Miss Rate) $= \frac{\sum FN}{\sum \text{условие положительно}}$	Истинно отрицательная пропорция, Специфичность, Селективность (True Negative Rate, сокр. *TNR*, Specificity, Selectivity) $= \frac{\sum TN}{\sum \text{условие отрицательно}}$	Отрицательное соотношение правдоподобия (Negative Likelihood Ratio, сокр. *LR*-) $= \frac{FNR}{TNR}$	Оценка F1 (F1 Score) $= \left(\frac{TPR^{-1} + PPV^{-1}}{2} \right)^{-1}$

Рисунок 4 — Матрица ошибок и полученные метрики эффективности классификации **[19]**

7.3 Уравненные шансы

Уравненные шансы (equalized odds) означают, что решения алгоритма не зависят от категории A при входных данных Y.

Предиктор $\hat{Y}$ удовлетворяет уравненным шансам в отношении категории A и результата Y, если $\hat{Y}$ и A независимы при условии Y

$$P(\hat{Y} = \hat{y} \mid Y = y, A = m) = P(\hat{Y} = \hat{y} \mid Y = y, A = n) \quad (1)$$

для всех значений Y, всех значений m, n из A.

Это означает, что истинно положительная пропорция (*TPR*) и ложноположительная пропорция (*FPR*) равны по демографическим категориям.

Стоит обратить внимание, что это определение позволяет моделям учитывать демографическую информацию. *TPR* равна 1 минус ложноотрицательная пропорция (*FNR*), следовательно *FNR* также равна по демографическим категориям. Сравнение *FNR* и ложноположительной пропорции (*FPR*) может помочь увидеть отношение между ложноотрицательным (*FN*) и ложноположительным (*FP*) результатом.

7.4 Равенство возможностей

Равенство возможностей означает, что решение алгоритма $\hat{Y} = 1$ не зависит от категории A, при входных данных $Y = 1$.

Бинарный предиктор $\hat{Y}$ удовлетворяет равенству возможностей в отношении к A и Y, если $\hat{Y} = 1$ и A независимы при условии $Y = 1$. Таким образом формально

$$P(\hat{Y} = 1 \mid Y = 1, A = m) = P(\hat{Y} = 1 \mid Y = 1, A = n) \quad (2)$$

для всех значений m, n из A.

Это означает, что истинно положительная пропорция (*TPR*) равна по демографической категории.

7.5 Демографический паритет

Статистический паритет (statistical parity) означает, что существуют равные коэффициенты предсказания между категориями. Демографический паритет (demographic parity), также известный, как групповая справедливость (group fairness), говорит о том, что существуют равные коэффициенты предсказания между демографическими категориями, например этнической принадлежностью. Демографический паритет, который является частным случаем статистического паритета, означает, что решение, такое как принятие или отклонение заявки на кредит — не зависит от демографического признака. Формально, дана демографическая переменная A

$$P(\hat{Y} = \hat{y} \mid A = m) = P(\hat{Y} = \hat{y} \mid A = n) \quad (3)$$

для всех значений m, n, которые может принимать A.

Паритет не учитывает случаи, когда выходное решение коррелирует с одной из оцениваемых групп или атрибутов (характеристик) ***[20]***, и нет гарантии, что сделанные прогнозы будут одинаково хороши для каждой категории.

7.6 Предсказательное равенство

Предсказательное равенство означает, что ложноположительная пропорция (*FPR*) равна по демографической категории. Формально

$$P(\hat{Y} = 1 \mid Y = 0, A = m) = P(\hat{Y} = 1 \mid Y = 0, A = n) \quad (4)$$

для всех значений m, n, которые может принимать A.

7.7 Другие метрики

Альтернативные метрики могут включать минимально-максимальную справедливость (minimax fairness) и справедливость Парето (Pareto fairness) ***[21]***.

8 Устранение нежелательной смещенности в течение всего жизненного цикла системы ИИ

8.1 Общие положения

Система ИИ или ИИ-сервис обычно проходит жизненный цикл от потребности бизнеса и стадии замысла через проектирование и разработку, верификацию и валидацию, до эксплуатации и вывода из эксплуатации. Жизненный цикл системы ИИ определен в ***[1]***[1)], разработанном в ПК 42 ***[22]***. Существуют различные способы определения жизненного цикла для конкретной услуги (сервиса) или продукта. В данном пункте описаны этапы жизненного цикла, значимые только для настоящего стандарта.

Во многих вариантах реализации (имплементации) систем ИИ части системы будут закупаться, а не разрабатываться одной и той же организацией. Учитывая это, различные части данного раздела будут применяться к различным контекстам реализации, и могут возникать соображения в отношении интеллектуальной собственности, прозрачности или коммерческие соображения, которые препятствуют выявлению и уменьшению смещенности.

Риски в цепочке поставок, связанные с нежелательной смещенностью, могут возникнуть, в частности, в случае отсутствия прозрачности исходного кода, моделей, происхождения обучающих данных или процессов разметки. Может быть полезным включение вопросов, связанных со смещенностью, в различные коммерческие соглашения.

8.2 Начальная стадия

8.2.1 Общие положения

Анализ системных требований является важным направлением деятельности по ослаблению нежелательной смещенности. Это этап, на котором анализируются внутренние и внешние требования, определяются заинтересованные стороны и участники (stakeholders), оцениваются цели системы. К этому этапу определены риски, связанные с системой, оценено воздействие на идентифицированные заинтересованные стороны и определены уровни вовлечения заинтересованных сторон.

Соображения и потенциальные требования, описанные в данном разделе, применимы не только к устранению нежелательной смещенности. Формальный анализ и более полный список соображений и других аспектов изложены в ***[23]*** и проектах ИСО/МЭК, разрабатываемых ПК 42 ***[22]***.

Все мероприятия по ослаблению смещенности осуществляются на основе политики, установленной руководящим органом, и посредством управленческой деятельности.

8.2.2 Внешние требования

Определение внешних требований как часть деятельности по системному анализу является обычной частью жизненного цикла разработки и закупки систем. В ходе этого процесса особое внимание может быть уделено следующим нормативным рамкам (frameworks):

- международным документам по правам человека, равенству (равноправию) и правам коренных народов, которые накладывают на организации обязательства по обеспечению определенных свобод, например, предоставление финансовых услуг без дискриминации;
- специальным законам и руководствам, касающимся предоставления технических решений, например регулирующих доступность программного обеспечения для пользователей с различными возможностями или регулирующих определенный сектор (например, ***[24]*** в США);
- законодательству о защите данных и конфиденциальности ***[25]***, которое может включать положения, касающиеся автоматизированного принятия решений. Это может быть наднациональное, национальное или региональное законодательство. На момент подготовки настоящего стандарта примеры законодательства о защите данных и конфиденциальности включают: ***[26]***, ***[27]*** и ***[28]***;
- законодательству о конкуренции и предпринимательстве.

Примеры возможных типов обязательств подотчетной организации:

- необходимость оценки рисков, которая может включать общественные (социальные) проблемы с точки зрения вовлеченных заинтересованных сторон;
- уведомление пользователей о том, что они подвергаются автоматизированному решению, требование получить ясно выраженное и однозначное согласие и предоставить неавтоматизированную альтернативу в случае отсутствия согласия;

[1)] Нормативная ссылка на международный стандарт ИСО/МЭК 22989 заменена на справочную ссылку из-за отсутствия гармонизированного с ним национального стандарта и его перевода на русский язык в Федеральном информационном фонде стандартов.

- обеспечение определенного уровня контролируемости или объяснимости решения для поддержки анализа конкретного решения или события;

- деятельность по количественной оценке или снижению рисков, например сбор метаданных об источниках данных для понимания их происхождения и качества ***[29]***;

- обеспечение значимого участия человека в процессе принятия решений;

- эквивалентное ценообразование и предоставление услуг для групп людей с определенными характеристиками.

Это может включать способность продемонстрировать, что равенство (равноправие) достигается на практике.

8.2.3 Внутренние требования

В дополнение к нормативным (регуляторным) требованиям многие другие факторы могут способствовать стремлению заинтересованной стороны ослабить смещенность, такие как:

- внутренние цели, стратегии и политика организации;

- моральные или культурные ценности;

- избежание общественных проблем (общественного осуждения, критики и т. п.) или репутационного ущерба.

В процессе анализа можно уделить особое внимание пяти конкретным областям: привлечение междисциплинарных экспертов, определение вовлеченных заинтересованных сторон, выбор источников данных, внешние изменения и спецификация (определение) критериев приемлемости, включая допустимые уровни смещенности.

8.2.4 Междисциплинарные эксперты

Хотя нежелательная смещенность является относительно новой проблемой в контексте технологий, она хорошо изучена в социальных науках. В рамках процесса анализа требований (и, более того, всего жизненного цикла системы) целесообразно предусмотреть доступную экспертизу, чтобы полностью снизить озабоченность общества по поводу смещенности (предвзятости) и принять во внимание различные перспективы. Такую экспертизу могут обеспечить:

- социологи и специалисты по этике;

- специалисты по данным и качеству;

- эксперты в области права и конфиденциальности данных;

- представители пользователей или группы внешних заинтересованных сторон.

Например, разработчики системы распознавания лиц могут придавать большое значение характеристике контура лица в своем проекте и упустить тот факт, что контур может быть (частично или полностью) закрыт у людей с определенными культурными или религиозными традициями. Достаточно разнообразная команда с большей вероятностью выявит такие ограничения в проектах, предполагаемых допущениях и наборах данных.

8.2.5 Определение заинтересованных сторон (стейкхолдеров)

Традиционный анализ требований включает в себя определение заинтересованных сторон ([stakeholders). Однако для того, чтобы соответствовать аспектам вышеупомянутой нормативной базы и должным образом смягчить общественные проблемы, это традиционное определение (идентификация) заинтересованных сторон может быть расширено и включать также тех, на кого прямо или косвенно влияет внедряемая система.

Исходя из типов данных, используемых для принятия автоматизированных решений, разработчики могут дополнительно разделить списки заинтересованных сторон на группы людей, которые поразному подвержены влиянию смещенности и предвзятости (предубеждений) в системе, обладают различными способностями в использовании системы или имеют различные уровни знаний и доступа. Важно рассмотреть, какие смещенности, негативный опыт или дискриминационные результаты могут возникнуть.

Это можно выяснить с помощью различных методов совместного проектирования или этнографических методов, такие методы предполагают активную работу и обсуждение с заинтересованными группами. Как правило, недостаточно зафиксировать на кого теоретически может быть оказано воздействие без непосредственного участия этих групп. Часто также недостаточно, чтобы один член заинтересованной группы (который, возможно, также работает в составе проектной группы) оценил, как эта группа подвергается воздействию. Группы не являются монолитным организмом, и один человек не всегда может адекватно представить весь спектр возможных точек зрения.

Выявление и привлечение заинтересованных сторон может быть включено в формальное описание и документацию по предполагаемым проблемам и потенциальным последствиям для заинте-

ресованных групп, как положительным, так и отрицательным. На этой основе можно получить более квалифицированные и количественные требования. Оценка воздействия на человека (human impact assessment), проводимая на более поздних этапах, может затем вернуться к этим проблемным областям и оценить, насколько успешно удалось смягчить их последствия.

8.2.6 Выбор и документирование источников данных

Выбор данных, используемых для формирования явных правил в экспертной системе, основанной на правилах, либо выбор данных, используемых для обучения моделей МО, является ответственным действием, которое оказывает значительное влияние на смещенность.

В случае с очевидными (явными) знаниями (explicit knowledge) важно учитывать когнитивную предвзятость человека, которая уже присутствует у тех, кто определяет эти знания. Когнитивная предвзятость может присутствовать в человеческих суждениях, которые затем кодифицируются в системе, основанной на правилах, а затем распространяются на весь срок службы системы в больших масштабах. Если принятие такой когнитивной предвзятости в одном (единичном) человеческом решении можно считать приемлемым, то распространение этой предвзятости через автоматизированное принятие решений может иметь гораздо большее влияние.

Статистические системы ИИ, которые обучаются на основе данных без явного определения знаний, подвержены многим рискам. Насколько это возможно, при сборе данных можно получить данные, справедливые для каждой группы, особенно в отношении результатов. Например, для снижения смещенности бинарного классификатора (binary classifier) собранные данные могут быть направлены на равное соотношение положительных и отрицательных обучающих примеров классификации для каждой заинтересованной группы.

Отдельные источники данных могут быть рассмотрены на предмет определения:

- полноты. Источник данных, который исключает определенные записи, поскольку не содержит одинаковых характеристик для всех записей, может дать неполную картину и привести к дефектам в процессе обучения. Общедоступные данные (например, из Интернета) вряд ли будут иметь эквивалентное распределение по группам людей;
- точности. Источник данных, содержащий неточные данные, будет распространять эти неточности и искажения на модель МО. Это может привести к общим проблемам с достоверностью (точностью), но эти проблемы также могут быть искажены (сдвинуты) в сторону определенных групп людей, например людей с худшей кредитной историей;
- процедуры сбора. Важно понимать историю происхождения данных, как они собираются, как вводятся и влияют ли эти процессы на полноту и точность. Можно учитывать местоположение (локализацию) и среду, из которой получены записи данных;
- своевременности. Частота сбора или обновления записей данных может иметь значение для обеспечения их точности. И наоборот, обновление может потребовать переоценки (пересмотра). Система, прошедшая аудит, может не соответствовать требованиям, особенно если обновления происходят в результате процесса, отличного от процесса по сбору первоначальных данных;
- последовательности. Последовательность, с которой определяются элементы входных данных (или элементы разметки), может иметь большое значение. Например, если человек классифицирует элементы, которые не имеют четкой границы между категориями, на основе одних и тех же данных могут быть получены различные категории или варианты разметки.

8.2.7 Внешние изменения

Целесообразно обратить внимание на то, какие изменения могут произойти в использовании разработанной или закупленной системы. Примеры включают:

- внедрение существующей системы в другой среде, включая других пользователей, целевые рынки и источники данных, может изменить риски, связанные с системой;
- со временем взаимосвязь между входами и выходами системы может измениться. Например, система, использующая модель МО для принятия решений на основе корреляций, установленных во время первоначального обучения модели МО, может пострадать, если эти корреляции со временем изменятся. Это известно как дрейф данных (data drift);
- варианты использования системы могут развиваться, как преднамеренно, так и органически, что требует переоценки рисков;
- общественные нормы со временем меняются ***(например, отношение к возрасту получения избирательного права, идеальной форме тела или курению)***[1)]. Смещенность в системах

[1)] Перечень примеров скорректирован для учета особенностей, характерных для Российской Федерации.

ИИ может быть пересмотрена и переосмыслена с учетом возникающих изменений (например, метрик, рисков, заинтересованных сторон или требований), и эти изменения могут быть учтены соответствующим образом.

8.2.8 Критерии приемлемости

Эффективные требования поддаются проверке, так как при их оценке можно определить, соответствует ли им система. Часто производительность системы ИИ (AI system performance) сравнивается с производительностью и способностями человека. Тем не менее, полезно иметь возможность указать производительность в статистической форме. Например, заинтересованные стороны могут указать предел для ложноположительных или ложноотрицательных решений в дополнение к общему показателю точности.

Предварительное определение приемлемости с точки зрения конкретных характеристик системы и степени их соответствия позволяет проводить эффективную оценку и принимать решения.

Критерии отказа могут быть нижней границей приемлемости, фактически устанавливая четкие границы приемлемых характеристик для модели. Если эти критерии не устанавливаются и не контролируются, система ИИ может дрейфовать (drift), что приведет к возникновению нежелательной смещенности, которая не будет замечена или исправлена. Процесс, связанный с выходом системы из строя (возникновение сбоев и погрешностей), также может быть тщательно рассмотрен в ходе проектирования для предотвращения экстремальных случаев возникновения нежелательной смещенности.

8.3 Проектирование и разработка

8.3.1 Общие положения

Модели сами по себе могут содержать нежелательную смещенность, если не принять меры по ее предотвращению. Предвзятость человека может быть закодирована в системах МО через неявные предположения, которые отразятся при проектировании. Таким образом, важно выявить и сделать явными неявные предположения.

В дополнение к содержанию данного пункта, инструменты с открытым исходным кодом, перечисленные в приложении А, могут помочь в решении вопросов смещенности в процессе проектирования и разработки.

8.3.2 Представление данных и разметка

8.3.2.1 Общие положения

Ключевым шагом в разработке системы МО является принятие решения о том, как наилучшим образом представить обучающие данные в виде признаков, интерпретируемых моделью. Это также называется разработкой или инжинирингом признаков (feature engineering), в 6.3 описаны некоторые типы смещенности данных, которые могут повлиять на этот процесс. Есть несколько зачастую неявных критериев, которые участвуют в этом процессе, включая критерии, по которым данные оцениваются как «хорошие» или «плохие» (например, можно ли оставить в наборе данных переэкспонированную фотографию). Эти критерии, касающиеся того, какие записи данных включаются в обучающие данные и какие признаки отбираются, можно сделать явными. Важно рассмотреть, как данные соотносятся с целью создания системы, процессом выбора характеристик, а также лицами, выбирающими характеристики, и их обоснованием (включая любые связанные с ними явно выраженные предположения). Важно оценить выбранные характеристики на предмет любых данных и когнитивной предвзятости человека, например, обратить внимание на отсутствующие значения признаков, неожиданные значения признаков или на перекос данных (data skew). Любой из этих факторов может указывать на то, что определенные группы или характеристики неточно представлены в данных.

Отсутствующие значения признаков могут быть результатом неявной смещенности в процессе сбора данных, которую можно выявить и устранить.

В алгоритмах глубокого обучения, где признаки создаются в процессе обучения, правильные метки имеют решающее значение. Аннотации, сделанные людьми для создания меток для данных, могут быть смещенными из-за когнитивной предвзятости человека или ошибок, возникающих из-за трудностей в самой задаче разметки. Важно убедиться в правильности меток, оценивая как специалистов по разметке, так и конечные размеченные данные. Кроме того, даже при правильных метках, типы меток, указанные специалистами по разметке, могут быть причиной нежелательной или необнаруженной смещенности итоговой модели.

8.3.2.2 Использование крауд-работников для разметки

Работники из числа разнообразных организаций и из числа пользователей/потребителей часто аннотируют и размечают данные, которые используются для контролируемого машинного обучения.

Ошибки или когнитивные предубеждения человека, проявляющиеся во время этого процесса, распространяются на обученную модель ***[30]***.

Если для разметки данных используются крауд-работники, может быть полезно изучить разнообразие и цели людей, аннотирующих данные, а также способы их стимулирования. Например, можно рассмотреть, как выглядит успех для разных работников, за что им платят (качество или количество) и компромиссы между временем, потраченным на выполнение задания, и удовольствием от задания, а также их культурное и социально-демографическое происхождение.

Разработчики могут составлять задания, учитывающие человеческие различия (и когнитивную предвзятость) при аннотировании, например, используя золотые тестовые вопросы с известными ответами или другие формы предварительного отбора участников.

Ясность инструкций, а также получение обратной связи от крауд-работников по потенциально сложным заданиям могут быть важны для снижения нежелательной смещенности. Человеческая изменчивость, включая доступность, мышечную память и когнитивную предвзятость при аннотировании, может быть учтена путем использования стандартного набора вопросов с известными ответами.

8.3.3 Обучение и настройка

8.3.3.1 Данные для обучения

Во многих случаях подготовка или создание сбалансированного набора данных может занять большую часть времени разработки. Кажущийся простым подход к снижению смещенности заключается в удалении соответствующих признаков, которые могут непосредственно отвечать за нежелательную смещенность. Например, в случае использования автоматического составления короткого списка кандидатов на основе информации из резюме примерами таких признаков могут быть этническая принадлежность, пол и возраст. В то же время к характеристикам, которые имеют отношение к данному сценарию использования, относятся опыт, навыки, квалификация, сертификация и профессиональная принадлежность. Хотя удаление этнической принадлежности, пола и возраста может показаться решением проблемы, другие признаки и элементы, выступающие в качестве косвенных переменных, могут косвенно отражать предвзятость. Например, такие характеристики, как приветствие или префикс (Mr/Ms/Mrs), или род занятий могут представлять собой косвенную переменную для пола. Таким образом, удаление только некоторых очевидных признаков, которые связаны с нежелательной смещенностью, не всегда приводит к ее ослаблению. Другие примеры косвенных переменных включают музыкальные вкусы и возраст, характер покупок и пол, почтовые индексы и расу, уровень дохода, семейное положение и пол, образование (какой университет или колледж окончил человек) и расу, вес или рост и пол и т. д.

Методы, основанные на данных, могут быть использованы для снижения смещенности в обучающих данных. Например, дополнительная (повторная) оценка данных может повысить вес выборки, которая соответствует поставленной цели. К таким методам относятся:

- методы отбора образцов для измерения репрезентативности выборки из различных источников, чтобы выявить и снизить смещенность;
- стратификация выборки для преодоления феномена редкости (rareness phenomenon). Такая выборка может быть использована путем увеличения относительной частоты для положительных случаев по сравнению с отрицательными. Это может быть сделано с помощью нескольких методов, включая методы синтетического перебора меньшинств (synthetic minority oversampling techniques, SMOTE) ***[31]***;
- тщательный отбор признаков в случаях, когда признаки выборки имеют сильную корреляцию со смещенностью, которую необходимо исключить (например, пол или цвет кожи).

Другой подход заключается в том, чтобы выяснить количество нежелательных смещений, присутствующих в данных, и убрать эти смещения из результата. Используя ряд шагов, можно определить вклад признака и относительную значимость каждого признака в предсказании модели. Затем можно компенсировать все влияние признака, вызвавшего смещение. Процесс определения относительной значимости признака может включать следующее:

- итеративная ортогональная проекция признаков (iterative orthogonal feature projection, IOFP). Учитывая входные и выходные данные модели МО, метод стремится создать рейтинг входных данных, который соответствует зависимости системы МО от каждого входного сигнала в процессе принятия решений и, таким образом, может обнаружить смещенность, связанную с определенными признаками ***[32]***;
- минимальная избыточность, максимальная релевантность (minimum redundancy, maximum relevance, MRMR). Отбор признаков определяет подмножества данных, которые релевантны используемым параметрам. Одна из схем заключается в выборе признаков, которые наиболее сильно коррелируют с классификационной переменной и при этом взаимно удалены друг от друга. Эта схема, назы-

ваемая отбором по принципу «минимальная избыточность — максимальная релевантность» (MRMR), оказалась более эффективной, чем другие способы отбора признаков ***[33]***;

- гребневая регрессия или регрессия LASSO (ridge or LASSO regression). Гребневая регрессия или регрессия LASSO — это методы линейной регрессии с регуляризацией для предотвращения избыточного перебора и подгонки к обучающим данным. Эти методы также используются для помощи в отборе признаков ***[34]***;

- случайный лес (random forest). Случайный лес — это подход, который объединяет несколько произвольно и случайно выбранных (рандомизированных) деревьев решений и агрегирует их прогнозы путем усреднения. Этот метод показал отличную производительность в условиях, когда количество переменных намного больше, чем количество наблюдений ***[35]***.

Обратите внимание, что хотя эти подходы используются для отбора признаков или их релевантности, они не всегда напрямую применимы для определения смещенности, присутствующей в данных.

8.3.3.2 Настройка

Для достижения различных целей были созданы алгоритмы ослабления смещенности (bias mitigation algorithms). Алгоритмы ослабления смещенности (также иногда называемые справедливыми алгоритмами — fair algorithms) можно классифицировать следующим образом:

- методы на основе данных, такие как увеличение выборки недопредставленных групп населения или использование синтетических данных;

- методы на основе модели, такие как добавление условий регуляризации или ограничений, которые усиливают цель во время оптимизации, или обучение представленности, чтобы скрыть или уменьшить влияние конкретной переменной;

- специальные пост-методы (post-hoc methods), такие как определение пороговых значений для принятия решений по конкретной группе на основе прогнозируемых результатов для выравнивания коэффициентов ложных срабатываний или других соответствующих метрик.

Примерами применяемых алгоритмов ослабления предвзятости являются:

- устранение разрозненного влияния: метод предварительной обработки, который редактирует значения, которые будут использоваться в качестве признаков, таким образом, чтобы уменьшить различия в отношении между группами;

- обнаружение и устранение индивидуальной смещенности: техника, которая создает новую модель МО для индивидуумов в неблагоприятной группе, получающих другое решение по сравнению с аналогичными индивидуумами в благоприятной группе. Иногда он может применять различные пороговые значения для положительной классификации в разных группах;

- разделенные классификаторы: метод обучения отдельного классификатора для каждой группы. Отдельные классификаторы можно эквивалентно рассматривать как единый классификатор, который разветвляется по групповому признаку;

- совместная функция потерь: метод учета паритета групп с помощью совместной функции потерь, которая ограничивает («наказывает») различия в статистике классификации между группами;

- трансферное обучение: метод ***[36]*** для ослабления проблем, связанных с малым объемом данных, для групп, в которых имеется меньшая совокупность данных.

8.3.4 Состязательные методы для ослабления смещенности

Одним из методов ослабления смещенности является включение состязательного блока в архитектуру модели ***[37]***. В этих методах «противник» предсказывает некоторое свойство или характеристику, определяющую группы, по отношению к которым желательна справедливость. Выходные данные модели, для которой необходимо ослабить смещенность, являются входными данными для модели противника. Обновление веса для этой модели затем изменяется таким образом, чтобы не только оптимизировать ее для выполнения задачи, но и уменьшить количество информации, которую она предоставляет противнику, полезной для его прогноза. Чистый эффект этой системы заключается в том, что система учится выполнять свою задачу способами, ортогональными к характеристикам, для которых смещенность нежелательна.

8.3.5 Нежелательная смещенность в системах, основанных на правилах

Разнообразие опыта и знаний разработчиков наряду с привлечением междисциплинарных экспертов (см. 8.2.4) может помочь снизить вероятность внесения нежелательной смещенности в разработку системы. Например, рассмотрим систему для автоматической идентификации потенциальных контрабандистов с жестко закодированными правилами, основанными на знаниях нескольких очень опытных экспертов. Если профильные эксперты в основном имеют опыт работы в определенной сфере контрабанды, использование системы в другой сфере может привести к непреднамеренному профи-

лированию определенных категорий людей. Это, в свою очередь, может привести к систематическому различию в отношении к этим классам по сравнению с другими классами в новом контексте. Обычным последствием такого сценария будет несбалансированная вероятность ошибочного распознавания человека (объектов) между различными когортами.

8.4 Верификация и валидация

8.4.1 Общие положения

Проверка и валидация недавно разработанной модели МО может выявить и снизить потенциальную нежелательную смещенность до этапа внедрения. При проверке и валидации обычно используется набор данных, полученный из источника данных, независимого от обучающего набора данных. Эта гарантия обобщаемости модели (model generalizability) также важна для защиты от любой нежелательной смещенности, скрытой в обучающем наборе данных. В целом, любые шаги, предпринятые во время обработки набора данных и обучения модели, полезно применять к валидации данных и процедур, если это возможно.

В то время как проверка систем МО проводится интенсивно с использованием обучающих и тестовых наборов данных (см. 8.3.3.2), она ограничивается проверкой результатов на основе отбора и вариации имеющихся данных. Система ИИ может быть оценена в конкретном контексте. Наличие отдельных команд, работающих над обучением и оценкой (это является обычной практикой при разработке программного обеспечения), также позволяет избежать влияния индивидуальной когнитивной предвзятости.

Исследование очевидных дефектов в модели может выявить, почему она не обеспечивает максимальную общую точность. Устранение этих дефектов может повысить общую точность. Наборы данных, недостаточно репрезентативные для определенных групп (см. 6.3.4), могут быть расширены дополнительными обучающими данными для повышения точности принятия решений и снижения смещенности результатов.

Тестирование программного обеспечения традиционно опирается на «совокупность знаний, используемых в качестве основы для разработки тестов и тестовых примеров» ***[38]***. Успех любой деятельности по эмпирическому тестированию обычно ограничен степенью, в которой окружающие требования или процессы управления рисками явно определили потенциальные нежелательные отклонения или источники нежелательных отклонений. Дополнительная информация об управлении рисками применительно к ИИ изложена в приложении Б.

Методы, описанные в данном подразделе, предназначены для использования в статистически значимом масштабе. Эти методы обычно измеряют чувствительность результата к подгруппе, не включенной явно в исходные данные (см. 8.3.3.1).

Смещенность в системах ИИ измеряется аналогично тому, как измеряются другие свойства, например, совокупная производительность (aggregate performance). Однако совокупные показатели производительности по всему тестовому набору не обязательно показывают, присутствует ли в модели нежелательная смещенность. Общая метрика в матрице ошибок (confusion matrix) (см. 7.2) может показаться хорошо работающей на всем множестве. Однако расчет точности и полноты (recall) на подмножествах демографически важных или определенных категорий часто может выявить смещенность, например, более низкую точность для одного идентифицированного пола по сравнению с другим или более низкую точность для определенной демографической группы. Эти различия в производительности, вероятно, указывают на то, что не выявленная смещенность присутствует на более ранних этапах процесса разработки. Например, определенная группа может быть недопредставлена (см. 6.3.4) в обучающих данных. Этот подраздел предназначен для применения при разработке новых систем, при внедрении существующих систем и при оценке того, сохраняется ли качество систем с течением времени. Изменение соотношения между ожидаемыми и фактическими входными данными может быть причиной для оценки. Оценка может также включать результаты внедрения системы для пользователей и сторонних наблюдателей (например, людей или объектов, которые случайно присутствуют, но не являются целью или объектом внедрения системы ИИ). Например, система, которая несправедлива по отношению к полу и этнической принадлежности независимо друг от друга, может быть справедлива по отношению к определенной комбинации этих двух факторов.

8.4.2 Статический анализ обучающих данных и подготовка данных

Анализ обучающих и операционных данных может выявить смещенность данных, например, описанную в 6.3. Методы кластеризации и визуализации обучающих данных, полученных характеристик

или результирующих прогнозов могут помочь обнаружить дисбаланс или потенциальную смещенность в обучающих данных или системе.

Эксперты могут определить профиль обучающих и эксплуатационных данных и проверить, соответствует ли разброс определенной переменной ожидаемому реальному набору данных. Примером может служить выявление того, что для обучения были использованы записи определенной возрастной группы, в то время как в реальных наборах данных ожидается другое распределение возрастов. Эта деятельность может быть направлена на проверку возможности смещенности отбора, смещенности выборки и смещенности охвата, но не может сделать это полноценно, поскольку эти процедуры будут ограничиваться знаниями специалиста по оценке.

Специалисты по оценке могут определить этапы процесса подготовки данных, на которых потенциально может возникнуть смещенность из-за «отсутствующих данных». Например, если определенные элементы данных недоступны последовательно во всем наборе исходных данных, инженеры могут применить эту информацию к оставшимся записям или удалить ее. Если отсутствие этого элемента данных коррелирует с определенными группами записей, это может привести к нежелательной смещенности, которая обычно не обнаруживается при тестировании модели.

8.4.3 Выборочная проверка меток

Риск некорректной разметки, описанный в 6.3.3, то есть неправильного указания человеком меток для набора входных данных, используемых затем для обучения модели, можно оценить путем выборочной проверки представленных меток.

Разметку, основанную на экспертной оценке, оценить сложнее. Можно провести двойную слепую проверку или рассмотреть возможность оценки несколькими экспертами, чтобы оценить качество первоначальной метки.

8.4.4 Тестирование внутренней валидности

Тестирование на внутреннюю валидность (достоверность) (internal validity testing) оценивает корреляцию между отдельными элементами входных данных и выходными данными системы. Затем тестирование на внутреннюю валидность проверяет, являются ли эти корреляции неблагоприятными в контексте конкретных требований или критериев приемки.

Этот процесс основан на том, что элементы данных, которые вызывают нежелательную смещенность, должны быть включены в область входных данных. Этот процесс может выявить смещенность в моделях и их взаимодействии, таких как экспрессивность (выраженность) модели, описанная в 6.4.7.2, нерепрезентативная выборка, описанная в 6.3.4, или проблемы обработки данных, описанные в 6.3.6.

Это может включать оценку в полностью интегрированной среде, чтобы выявить любую нежелательную смещенность при сборе или подготовке данных, используемых при разработке системы ИИ. Интеграция также может выявить нерепрезентативную выборку. Например, данные, собранные в интегрированной среде, могут иметь различные характеристики, такие как уровень освещения или частота обновления датчиков. Эти изменения могут повлиять на входные данные для системы ИИ.

8.4.5 Тестирование внешней валидности

Проверка внешней валидности (external validity testing) может включать в себя переоценку предыдущих наблюдений с использованием внешних источников данных. Это полезный метод, поскольку он позволяет выявить многие типы нежелательной смещенности, описанные в настоящем стандарте, включая непрямую (косвенную) смещенность (indirect bias). Аспект входных данных, к которому относится косвенная предвзятость, не содержится в явном виде в характеристиках, а является производной второго порядка ***[39]***.

Например, некоторые сообщения СМИ о смещенности (предвзятости) ИИ ***[40]*** касались исследований, в которых результаты модели соотносились с данными переписи населения или почтовыми индексами, чтобы проиллюстрировать неравенство результатов.

Проверка внешней валидности также может включать интеграцию новых входных данных и проверку того, что результаты соответствуют результатам внутренней проверки валидности.

Проверка внешней валидности особенно важна для косвенной смещенности, вносимой прокси-переменными (proxy variables). Если разработчик модели пытается уменьшить смещенность, просто удаляя демографическую информацию из исходных данных, нежелательная смещенность, скорее всего, все равно будет существовать через прокси-переменные. Например, модель может укоренить (закрепить) «дальтонический» расизм ***[41]***, ***[42]***, социологическую концепцию, которая описывает, как утверждения о том, что человек не «видит» цвет кожи, препятствуют пониманию и решению проблемы расового неравенства в обществе. Чтобы избежать такого результата, проверка внешней валидности может фактически включать демографические данные, первоначально исключенные, или учитывать

исследование влияния прокси-переменной. Можно провести более глубокое исследование, чтобы понять, почему такие прокси-переменные существуют и можно ли достичь цели без них. Проверка внешней валидности может включать качественные данные, демонстрирующие разное влияние одной и той же классификации. Например, если определенная модель используется для идентификации людей перед посадкой в самолет, эмоциональный ущерб от ложноотрицательного результата может быть больше для тех групп, которые стереотипно считаются вероятными «террористами», чем для других групп. В этом контексте интеграция входных наборов данных с дополнительными значениями может быть полезной для правильной оценки системы на предмет нежелательной смещенности.

8.4.6 Тестирование с пользователями

Тестирование с различными категориями конечных пользователей может быть полезным, когда взаимодействие пользователя с системой влияет на результаты и прогнозы, при этом учитывается принадлежность пользователя к той или иной группе.

Оценка пользовательского опыта в реальных сценариях по широкому спектру категорий пользователей, сценариев использования и контекстов использования является полезной методикой для выявления нежелательной смещенности во взаимодействии моделей (см. 6.4.7), проблем с обработкой данных (см. 6.3.6) и проблем с разметкой данных (см. 6.3.3).

8.4.7 Исследовательское тестирование

Разработчики системы могут организовать команду разнообразных доверенных тестировщиков, которые могут выступать в роли противников для тестирования системы и включать различные потенциально опасные входные данные в модульные или функциональные тесты. Это может помочь в выявлении непредвиденных вариантов смещенности системы, особенно если в пул тестировщиков входят представители групп, на которые может влиять система и которые могут быть ее конечными пользователями.

8.5 Внедрение

8.5.1 Общие положения

После внедрения системы ИИ важно обеспечить надлежащее обучение и поддержку пользователей для эффективного использования продукта. Это включает в себя руководство для разработчиков системы, в котором поясняется что представляет собой соответствующая предусмотренная (надлежащая) и ненадлежащая эксплуатация системы ИИ. Например, система отслеживания внимания может восприниматься как неуместная, если она используется в образовательной системе для мониторинга поведения учащихся, но все может измениться, если та же система используется в качестве исследовательского инструмента в психологическом эксперименте.

Внедренные системы могут также включать руководство для конечных пользователей. Например, желательно, чтобы персонал по набору кадров, использующий систему рекомендаций по найму, понимал возможности и ограничения системы. Как разработчики ИИ, так и конечные пользователи могут быть осведомлены об известных областях нежелательной смещенности. Этого можно достичь с помощью инструмента прозрачности (см. 8.5.3), который содержит информацию о данных, на которых обучалась модель, о распределении для популяций ее ложноположительных и ложноотрицательных ошибок и другую сопутствующую информацию.

Субъекты данных — люди, к которым относятся обучающие данные, — не обязательно являются пользователями системы. Они не нуждаются в обучении, но их можно проинформировать о любой смещенности в системе, которые могут повлиять на них, на языке, соответствующем контексту. Недостатки в обучении или поддержке могут привести к дополнительной смещенности, которую трудно обнаружить на более ранней стадии

8.5.2 Постоянный мониторинг и валидация

Модели могут терять эффективность (результативность) (performance) с течением времени. Снижение эффективности может быть связано с изменениями в окружающем мире, такими как общественные тенденции, практика и нормы, появление новых моделей поведения, изменение состава входных данных и изменение требований. Кроме того, система может быть смещенной по отношению к исторической позиции ***[20]***.

Текущую эффективность, используя методы, описанные в 8.4, можно отслеживать, когда система введена в эксплуатацию. Это включает в себя проверку производительности системы, например результатов с выбросами, с использованием визуального исследования данных и методов оценки смещенности и справедливости, в том числе с применением автоматизированных средств. Более подроб-

ную информацию об измерении см. в разделе 7. Если имеются признаки нежелательной смещенности, систему можно переобучить или перепроектировать. Мониторинг является известной процедурой во многих отраслях, использующих в своих процессах автоматизированное принятие решений. Например, в банковской сфере разрабатываются и внедряются модели оценочных карт вместе с утвержденными процессами мониторинга. Процессы мониторинга могут применяться не только к точности и эффективности моделей (или систем), но также могут использоваться для выявления и отслеживания нежелательной смещенности в системах или моделях.

8.5.3 Инструменты прозрачности

Чтобы уточнить предполагаемые случаи использования моделей МО и минимизировать их применение в контекстах, для которых они не вполне подходят, реализованные модели могут снабжаться документацией с подробным описанием их эксплуатационных характеристик. Инструменты прозрачности моделей могут обеспечить основу для прозрачной отчетности о происхождении, использовании и справедливой оценке моделей МО. Документация к моделям может включать:

- качественную информацию, такую как этические соображения, данные о целевых пользователях, примеры эксплуатации;
- количественную информацию, состоящую из дезагрегированной (распределение по различным целевым подгруппам) и межсекторальной (включая оценку по нескольким подгруппам в сочетании, например, по этнической принадлежности и полу) оценки модели. Дополнительную информацию о метриках см. в разделе 7;
- информация о данных, если это возможно, которая может быть формализована в виде инструмента прозрачности данных.

Полезность и точность инструмента прозрачности зависит от добросовестности создателя (создателей) самого инструмента и может храниться в виде документации или метаданных, связанных с каждой моделью. Карточки моделей ***[43]*** — это один инструмент прозрачности среди многих других, которые могут включать, например, алгоритмический аудит третьими сторонами (как количественный, так и качественный), «состязательное тестирование» техническими и нетехническими аналитиками и более комплексные механизмы обратной связи с пользователями (см. 8.4.6).

При использовании (или предоставлении) сторонних моделей и приложений, также известных как «ИИ как услуга» (AI as a service) или «МО как услуга» (ML as a service) для повышения прозрачности и доверия к таким предложениям можно использовать специальные листы контроля «FactSheets» ***[44]***.

Приложение А
(справочное)

Примеры смещенности

А.1 Пример 1

Рассмотрим алгоритм, который используется в процессе подачи заявки на кредит, чтобы сделать прогноз о том, представляет ли заявитель приемлемый риск или нет.

Требуемая система ИИ будет правильно предсказывать, представляет ли заявка приемлемый риск, не способствуя систематическому исключению определенных групп.

Гипотетическим примером неправильного прогноза может быть отклонение заявки на кредит от кандидата с «приемлемым риском». В таком сценарии система ИИ автоматизирует существующий процесс, в котором кредитные специалисты определяют, представляют ли заявки приемлемый риск. В течение многих лет этот процесс, управляемый человеком, может привести к тому, что 25 % заявок от иммигрантов будут отклонены. Перед тестированием алгоритма на смещенность было решено, что, учитывая существующий уровень предвзятости в обществе и то, что ложноотрицательные результаты (т. е. отказ заявителям, которые на самом деле являются кредитоспособными) считаются более важными, чем ложноположительные результаты, демографического паритета и равенства возможностей в пределах 2 % будет достаточно, чтобы признать модель прогнозирования «несмещенной» («непредвзятой»).

В первой версии система ИИ, находящаяся в эксплуатации и обученная на всех предыдущих кредитных заявках, отклонила 20 % заявок иммигрантов, но только 10 % заявок от людей, не являющихся иммигрантами. Система ИИ обучалась на основе решений человека и имитирует их, при этом, хотя алгоритм работает лучше, чем процесс, управляемый человеком, он не проходит тест на демографический паритет и отклоняется как неприемлемый. После удаления чувствительных факторов из обучающих данных новая модель показала результаты, приведенные в таблицах А.1 и А.2 в данном приложении. В этих таблицах положительное условие представляет кредитоспособного заявителя, а отрицательное условие — кредитный риск.

Таблица А.1 — Матрица ошибок для заявлений иммигрантов в примере 1

		Истинные условия (True condition)		
	Общая популяция (Total popula tion)	Условие позитивно (Condition positive)	Условие негативно (Conditio negative)	Общий прогноз, предсказание (Total prediction)
Условие прогноза, предсказания (Predicted condition)	Позитивный прогноз (Prediction positive)	88	0	88
	Негативный прогноз (Prediction negative)	2	10	12
	Общее условие (Total condition)	90	10	100
		TPR = 0,98	*TNR* = 1,00	*ACC* = 0,98
		FPR = 0,00	*FNR* = 0,02	

Таблица А.2 — Матрица ошибок для заявлений неиммигрантов в примере 1

		Истинные условия (True condition)		
	Общая популяция (Total popula tion)	Условие позитивно (Condition positive)	Условие негативно (Condition negative)	Общий прогноз, предсказание (Total prediction)
Условие прогноза, предсказания (Predicted condition)	Позитивный прогноз (Prediction positive)	89	1	90
	Негативный прогноз (Prediction negative)	1	9	10
	Общее условие (Total condition)	90	10	100
		TPR = 0,99	*TNR* = 0,90	*ACC* = 0,98
		FPR = 0,10	*FNR* = 0,01	

Поскольку уровень отказов для иммигрантов находится в пределах 2 % от уровня отказов для неиммигрантов, действует демографический паритет. А поскольку *TPR* для иммигрантов находится в пределах 2 % от *TPR* для неиммигрантов, также имеет место равенство возможностей (equality of opportunity holds). Таким образом, в соответствии с критериями смещенности, установленными до начала тестирования, новые модели можно считать несмещенными (непредвзятыми) и готовыми к применению.

А.2 Пример 2

В другом гипотетическом сценарии предприятие хочет применить ИИ, чтобы войти в новую сферу бизнеса на основе анонимизированных данных, которые оно считает полезными, но которые не дают полного представления о кредитоспособности. Предприятие решает использовать демографический паритет в качестве метрики для определения степени несправедливой смещенности в обученной модели, установив порог в 1 % разницы для принятия модели. Испытания системы ИИ показывают, что она отклоняет 30 % заявок — независимо от иммиграционного статуса. Несмотря на то, что система ИИ отклоняет слишком много заявлений, это не связано с какой-либо конкретной группой; в соответствии с их выбором метрики и порога, модель является несмещенной. Бизнес решает, что высокий общий коэффициент ошибок — это приемлемый риск, на который можно пойти, учитывая, что предельные затраты на вхождение в это рыночное пространство низки, и это может стоить цены автоматизации. Однако это решение может быть весьма противоречивым. Например, в условиях, приведенных в таблицах А.3 и А.4, модель действительно удовлетворяет демографическому паритету, но не проходит все остальные тесты алгоритмической справедливости, определенные в разделе 7. Кроме того, она заметно хуже по точности для иммигрантов по сравнению с неиммигрантами, что оставляет организацию не защищенной (открытой) для обвинений в дискриминации.

Таблица А.3 — Матрица ошибок для заявлений иммигрантов в примере 2

		Истинные условия (True condition)		
	Общая популяция (Total popula tion)	Условие позитивно (Condition positive)	Условие негативно (Condition negative)	Общий прогноз, предсказание (Total prediction)
Условие прогноза, предсказания (Predicted condition)	Позитивный прогноз (Prediction positive)	65	5	70
	Негативный прогноз (Prediction negative)	15	15	30
	Общее условие (Total condition)	80	20	100
		TPR = 0,81	*TNR* = 0,75	*ACC* = 0,80
		FPR = 0,25	*FNR* = 0,19	

Таблица А.4 — Матрица ошибок для заявлений неиммигрантов в примере 2

		Истинные условия (True condition)		
	Общая популяция (Total popula tion)	Условие позитивно (Condition positive)	Условие негативно (Conditio negative)	Общий прогноз, предсказание (Total prediction)
Условие прогноза, предсказания (Predicted condition)	Позитивный прогноз (Prediction positive)	65	5	70
	Негативный прогноз (Prediction negative)	5	25	30
	Общее условие (Total condition)	70	30	100
		TPR = 0,93	*TNR* = 0,83	*ACC* = 0,90
		FPR = 0,17	*FNR* = 0,07	

Приложение В
(справочное)

Инструменты с открытым исходным кодом

В.1 Общие положения

Инструменты с открытым исходным кодом, включая перечисленные в данном приложении, доступны для аудита смещенности и объяснения результатов работы системы ИИ. Приведенный ниже список не является исчерпывающим, имеются и разрабатываются и другие инструменты. Эти инструменты приведены в качестве примеров.

В.2 Инструменты

ELI5: пакет Python ***[45]***, который помогает отлаживать классификаторы МО и объяснять их предсказания (прогнозы). Он обеспечивает поддержку ряда пакетов, таких как scikit-learn, Keras, XGBoost.

FairML: инструментарий ***[46]***, написанный на Python, для аудита моделей МО на предмет справедливости и смещенности. Он помогает количественно оценить значимость входов модели. Он использует четыре алгоритма ранжирования входов для количественной оценки относительной предсказательной зависимости модели от ее входов.

Google What-If Tool (WIT): плагин для тензорной доски ***[47]*** для понимания классификации «черного ящика» или регрессионной модели МО.

LIME: проект с открытым исходным кодом ***[48]***, направленный на объяснение и интерпретацию того, как работают модели МО. LIME поддерживает широкий спектр моделей МО. Он способен интерпретировать текстовую классификацию, многоклассовую классификацию, классификацию изображений и регрессионные модели.

AI Fairness 360: библиотека с открытым исходным кодом ***[49]***, которая обнаруживает и снижает смещенность в моделях МО с помощью набора алгоритмов устранения смещенности.

Fairlearn: инструментарий с открытым исходным кодом ***[50]***, состоящий из двух компонентов: интерактивной панели визуализации и алгоритмов устранения несправедливости. Эти компоненты предназначены для оценки и повышения справедливости систем ИИ при поиске компромиссов между справедливостью и эффективностью модели.

Skater: унифицированный фреймворк ***[51]***, позволяющий интерпретировать модели всех форм. Библиотека Python с открытым исходным кодом, которая помогает понять изученные структуры модели «черного ящика» как глобально (вывод на основе полного набора данных), так и локально (вывод об отдельном предсказании).

Shapley Additive exPlanations (SHAP): инструмент ***[52]***, который может объяснить выходные результаты любой модели МО, соединяя теорию игр с локальным пояснением. Он использует визуализацию для объяснения моделей.

FairTest: инструмент с открытым исходным кодом, который «позволяет разработчикам или аудиторским организациям обнаруживать и тестировать на наличие необоснованных ассоциаций между результатами алгоритма и определенными подгруппами пользователей, идентифицированными по защищенным признакам» ***[53]***.

Themis: подход к тестированию и инструмент для измерения дискриминации в программной системе ***[54]***.

Приложение С
(справочное)

Пример карты соотношений (*см. [55]*)

С.1 Общие положения

В настоящем приложении показано, как общественные интересы и взгляды, описанные в ***[55]***, соотносятся (коррелируют) с отношением к смещенности в системах ИИ, как описано в настоящем стандарте. Связь между дополнительными общественными проблемами и расширенным списком «уязвимостей» ИИ может быть показана таким же образом с использованием методологии управления рисками.

В таблицах данного приложения используются хорошо известные примеры из ***[56]***, чтобы продемонстрировать как идентификация целей, связанных с ними рисков и их возможного устранения происходит на различных общественных и организационных слоях (или уровнях). Примеры показывают, что по мере распространения управления рисками по уровням, идентифицированные средства управления верхнего уровня часто становятся целями этого уровня (или уровней).

[55] определяет семь «основных тем» с различными «вопросами» под каждой из них и описывает, как они могут быть приняты во внимание руководящими органами организаций.

С.2 Уровень общества

В качестве примера в таблице С.1 показано, как несколько вопросов, относящихся к основной проблеме «прав человека», могут быть рассмотрены или пересмотрены в эпоху ИИ.

Таблица С.1 — Уровень общества (например, ***[55]***)

Цель	Источник риска	Стратегии ослабления смещенности (Bias treatment strategies)
Предмет исследования: Права человека **Тема 5:** Дискриминация и уязвимые группы населения **Тема 7:** Экономические, социальные и культурные права	- Дискриминация по признаку этнической принадлежности, пола и т. д. - Непредоставление равного доступа к экономическим, социальным и культурным возможностям.	- Обеспечить обратную связь с ответственными за разработку политики в отношении использования ИИ. - Независимо тестировать или проверять системы ИИ на справедливость результатов. - Сделать информацию о системах ИИ прозрачной для всех заинтересованных сторон. - Привлекать заинтересованные стороны из затронутых групп к разработке, тестированию и оценке системы, чтобы помочь выявить и снизить смещенность.

С.3 Уровень управления организацией

Как отмечается в ***[55]***, работа, направленная на решение проблем общества, является одним из основных направлений деятельности руководящего органа организации. В таблице С.2 приведены примеры стратегий работы, которые доступны на разных уровнях управления.

Таблица С.2 — Уровень руководства организации

Цель	Источник риска	Стратегии ослабления смещенности (Bias treatment strategies)
Справедливость (Fairness)	- Системы ИИ могут быть использованы противозаконно или неэтично. - Ограничения технологий ИИ или МО.	- Создать совет по этике. - Определить и интегрировать юридические требования, связанные с смещенностью.
Прозрачность (Transparency)	- ИИ или МО предполагают сложные цепочки создания стоимости. - Масштабируемые модели МО непрозрачны.	- Определить процедуры межорганизационного управления для достижения прозрачности МО, как указано в 8.5.3.
Комплексная вовлеченность[1)], инклюзивность (Inclusiveness)	- Не учитываются права или ожидания соответствующих заинтересованных сторон.	- Продвигать широкую, комплексную вовлеченность специалистов при разработке систем как один из принципов организации внутри и вне организации, как рассмотрено в 8.2. - Привлекать работников или их представителей.

[1)] Добавлен синоним для учета особенностей, характерных для Российской Федерации.

С.4 Уровень менеджмента организации

В таблице С.3 приведены примеры возможных мер по снижению воздействия, методов устранения и контроля, чтобы показать, как смещенность может рассматриваться и устраняться на различных уровнях (или слоях) организации.

Таблица С.3 — Уровень менеджмента организации

Цель	Источник риска	Стратегии ослабления смещенности (Bias treatment strategies)
Справедливость (Fairness)	- Системы ИИ, содержащие смещенность или проявляющие несправедливость.	- Сбор и документирование (создание каталога) технических методов измерения и уменьшения смещения.
Прозрачность (Transparency)	- Проекты, работающие с ИИ-системами, могут по-разному понимать, что такое «прозрачность».	- Создать образец (шаблон) прозрачности, который будет использоваться для систем ИИ (т. е. продуктов или услуг). Например, реестры ИИ городов Амстердам ***[57]*** и Хельсинки ***[58]***. - Создать каталог подходов к объяснимости.
Комплексная вовлеченность[1)], инклюзивность (Inclusiveness)	- Заинтересованные стороны могут быть не замечены (игнорированы). - Заинтересованные стороны могут быть недоступны.	- Создать межинституциональную группу, объединяющую академические, общественные и другие организации по вопросам, связанным с ИИ. - Привлекать работников или их представителей.

С.5 Уровень проекта в сфере ИИ

В таблице С.4 приведены примеры стратегий по ослаблению смещенности, которые могут применяться на уровне проекта.

Таблица С.4 — Уровень проекта

Цель	Источник риска	Стратегии ослабления смещенности (Bias treatment strategies)
Справедливость (Fairness)	- Этические дилеммы. - Конфликт интересов заинтересованных сторон. - Противоречивые законы или регулирующие документы. - Отсутствие осведомленности или знаний.	- Выявлять новые этические дилеммы и, при необходимости, передавать их по цепочке управления и менеджмента, например в совет по этике (С.3). - Привлекать регулирующие органы и общественные группы как на уровне управления, так и на уровне требований. - Прекратить проект. - Использовать технические и профессиональные методы (см. 8.3) для устранения смещенности. - Проводить занятия или создать возможности для обучения этическим аспектам.
Прозрачность (Transparency)	- Проект выполняется в сжатые сроки по графику. - Проект является конфиденциальным. - Проект требует использования высокомасштабируемых, но нелегко понимаемых нейронных сетей.	- Заполнить шаблон прозрачности, как описано в таблице С.3. - Использовать интерпретируемые модели. - Предоставить инструменты объяснимости.
Комплексная вовлеченность[1)], инклюзивность (Inclusiveness)	- Проект выполняется в сжатые сроки. - Проект является конфиденциальным. - Отсутствие разнообразия.	- Выявлять и привлекать соответствующие заинтересованные стороны на каждом этапе жизненного цикла проекта или системы ИИ. - Обеспечить широкое и разнообразное представительство среди проектной команды и тестировщиков системы.

[1)] Добавлен синоним для учета особенностей, характерных для Российской Федерации.

С.6 Примеры стратегий рассмотрения и ослабления смещенности

В таблице С.5 приведены примеры стратегий работы со смещенностью из настоящего стандарта, которые могут быть применены.

Таблица С.5 — Примеры стратегий деятельности в области предвзятости

Цель	Источник риска	Стратегии ослабления смещенности (Bias treatment strategies)
Снижение смещенности требований (requirements bias)	- Когнитивная предвзятость человека. - Упущение важных характеристик. - Контекстуальные (например, географические) допущения.	- Привлекать соответствующих экспертов и заинтересованные стороны (см. «***Комплексная вовлеченность***[1)], инклюзивность»).
Снижение смещенности данных (data bias)	- Нерепрезентативная выборка. - Определение и создание меток (разметка). - Упущение важных характеристик. - Непреднамеренные изменения во время предварительных процессов или процессов пост-обработки. - Избыточные кодирование - Исторические и общественные предубеждения. - Смещенность при сборе данных.	- Определить возможные источники смещенности. - Оценить разметку и специалистов, которые этим занимаются. - Измерить возможную смещенность.
Снижение предвзятости модели (model bias)	- Особенности модели, разработанной вручную. - Информативность. - Взаимодействие между моделями. - Экспрессивность.	- Тестирование. - Оценка и измерение. - Настройка.

[1)] Добавлен синоним для учета особенностей, характерных для Российской Федерации.

Приложение ДА
(справочное)

Сведения о соответствии ссылочных национальных стандартов международным стандартам, использованным в качестве ссылочных в примененном международном документе

Таблица ДА.1

Обозначение ссылочного национального стандарта	Степень соответствия	Обозначение и наименование ссылочного международного стандарта
ПНСТ 838—2023 (ИСО/МЭК 23053:2022)	MOD	ISO/IEC 23053:2022 «Экосистема разработки систем искусственного интеллекта (ИИ) с использованием машинного обучения (МО)»

Примечание — В настоящей таблице использовано следующее условное обозначение степени соответствия стандарта:

- MOD — модифицированный стандарт.

Библиография

[1][1)]	***ИСО/МЭК 22989***	***Информационные технологии. Искусственный интеллект. Концепции искусственного интеллекта и терминология (Information technology — Artificial intelligence — Artificial intelligence concepts and terminology)***
[2]	ИСО 3534-1:2006	Статистика. Словарь и условные обозначения. Часть 1. Общие статистические термины и термины, используемые в теории вероятностей (Statistics — Vocabulary and symbols — Part 1: General statistical terms and terms used in probability)
[3]	ИСО/МЭК 2382:2015	Информационная технология. Словарь (Information technology — Vocabulary)
[4]	ISO/IEC/IEEE 15288:2015	Системная и программная инженерия. Процессы жизненного цикла системы (Systems and software engineering — System life cycle processes)
[5]	ИСО 20501:2019	Керамика тонкая (высококачественная керамика, высококачественная техническая керамика). Статистические данные о прочности по Вейбуллу [Fine ceramics (advanced ceramics, advanced technical ceramics) — Weibull statistics for strength data]

[6] KLEINBERG J., MULLAINATHANS & RAGHAVAN M. Inherent Trade-Offs in the Fair Determination of Risk Scores. 67(2017) pp: 43:1-43:23

[7] CALISKAN A., BRYSON J.J., and NARAYANAN A. Semantics derived automatically from language corpora contain human-like biases. Science, 356:183—186, 2017

[8] GILVOICH T., & GRIFFIN Dale W. 2010. Judgement and decision-making. In S. T. Fiske, D.T. Gilbert and G. Lindzey (Eds.). Handbook of Social Psychology, Fifth Edition, (Vol 1, pp. 542-589)

[9] KAHNEMAN D., FARRAR, STRAUS and GIROUX 2013. Thinking, Fast and Slow

[10] Machine Bias — ProPublica, 2016. [online]. [Accessed 23 September 2020]. Available from: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[11] TRENTA A. ISO/IEC 25000 quality measures for A.I.: a geometrical approach in Proceedings APSEC IWESQ 2020 (http://ceur-ws.org/Vol-2800/, ISSN 1613-0073)

[12] SURESH H., and GUTTAG J. V. 2020. A Framework for Understanding Unintended Consequences of Machine Learning. arXiv:1901.10002 [cs, stat] [online]. 17 February 2020. [Accessed 22 February 2020]. Available from: https://arxiv.org/abs/1901.10002

[13] VERMA S., and RUBIN J. 2018. Fairness definitions explained. In: Proceedings of the International Workshop on Software Fairness — FairWare '18 [online]. Gothenburg, Sweden: ACM Press. 2018. p. 1—7. [Accessed 6 May 2019]. ISBN 978-1-4503-5746-3. Available from: https://dl.acm.org/citation.cfm?doid=3194770.3194776

[14] MITCHELL S., POTASH E., BAROCAS S., D'AMOUR A. and LUM K. 2020. Prediction-Based Decisions and Fairness: A Catalogue of Choices, Assumptions and Definitions. arXiv:1811.07867 [stat] [online]. 24 April 2020. [Accessed 30 September 2020]. Available from: https://arxiv.org/abs/1811.07867

[15] SALEIRO P., KUESTER B., HINKSON L., LONDON J., STEVENS A., ANISFELD A., RODOLFA K.T., and GHANI R. 2019. Aequitas: A Bias and Fairness Audit Toolkit. arXiv:1811.05577 [cs] [online]. 29 April 2019. [Accessed 30 September 2020]. Available from: https://arxiv.org/abs/1811.05577

[16] GAJANE P., and PECHENIZKIY M. 2018. On Formalizing Fairness in Prediction with Machine Learning. arXiv:1710.03184 [cs, stat] [online]. 28 May 2018. [Accessed 30 October 2020]. Available from: https://arxiv.org/abs/1710.03184
arXiv: 1710.03184

[17] AGARWAL A., DUDÍK M., and WU Z.S. 2019. Fair Regression: Quantitative Definitions and Reduction-based Algorithms. arXiv:1905.12843 [cs, stat] [online]. 29 May 2019. [Accessed 30 October 2020]. Available from: https://arxiv.org/abs/1905.12843 arXiv: 1905.12843

[18] THARWAT A. 2020. Classification assessment methods. Applied Computing and Informatics [online]. 3 August 2020. Vol. ahead-of-print, no. ahead-of-print. [Accessed 30 September 2020]. DOI 10.1016/j.aci.2018.08.003. Available from: https://www.emerald.com/insight/content/doi/10.1016/j.aci.2018.08.003/full/html

[19] Confusion matrix, 2020. Wikipedia [online]. [Accessed 30 October 2020]. Available from: https:// en.wikipedia.org/w/index.php?title=Confusion_matrix&oldid=980584181 Page

[1)] Нормативная ссылка на международный стандарт ИСО/МЭК 22989 заменена на справочную ссылку из-за отсутствия гармонизированного с ним национального стандарта и его перевода на русский язык в Федеральном информационном фонде стандартов.

[20] DWORK C., HARDT M., PITASSI T., REINGOLD O., and ZEMEL R. 2011. Fairness Through Awareness. arXiv:1104.3913 [cs] [online]. 28 November 2011. [Accessed 22 February 2020]. Available from: https://arxiv.org/abs/1104.3913

[21] MARTINEZ N., BERTRAN M., SAPIRO G. 2020. Minimax Pareto Fairness: A Multi Objective Perspective. Available from: http://proceedings.mlr.press/v119/martinez20a/martinez20a.pdf

[22] ИСО/МЭК JTC 1/SC 42, Artificial intelligence, [no date]. [online]. [Accessed 30 October 2020], Available from: https://www.iso.org/committee/6794475/x/catalogue/p/0/u/1/w/0/d/0

[23] ИСО/МЭК 38507:2022 Информационные технологии. Стратегическое управление ИТ. Последствия влияния стратегического управления при использовании искусственного интеллекта организациями (Information technology — Governance of IT — Governance implications of the use of artificial intelligence by organizations)

[24] UNITED STATES, Health Insurance Portability and Accountability Act, 1996

[25] CHAUDHURI A., SMITH A.L., GARDNER A., GU L., SALEM M.B., and LÉVESQUE M. 2019. Regulatory frameworks relating to data privacy and algorithmic decision-making in the context of emerging standards on algorithmic bias

[26] UNITED STATES, California Consumer Privacy Act. AB-375 Privacy: personal information: businesses, 2018

[27] JAPAN, Act on the Protection of Personal Information 2016

[28] EUROPEAN PARLIAMENT. REGULATION (EU) 2016/679 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on the protection of natural persons with regard to the processing of personal data and on the free movement of such data and repealing Directive 95/46/EC (General Data Protection Regulation). 2016

[29] GEBRU T., MORGENSTERN J., VECCHIONE B., VAUGHAN J.W., WALLACH H., DAUMÉ III H. and CRAWFORD K. 2018. Datasheets for Datasets. arXiv:1803.09010 [cs] [online]. 23 March 2018. [Accessed 10 October 2019]. Available from: https://arxiv.org/abs/1803.09010

[30] STEED R. & CALISKAN A. Image Representations Learned With Unsupervised Pre-Training Contain Human-like Biases. ACM Conference on Fairness Accountability and Transparency (FAccT '21), 2021, Virtual Event, Canada. ACM, New York, NY, USA, pp. 701-713. Available from: https://doi.org/10.1145/3442188.3445932

[31] CHAWLA N. V., BOWYER K. W., HALL L. O., and KEGELMEYER W. P. 2002. SMOTE: Synthetic Minority Oversampling Technique. Journal of Artificial Intelligence Research. 1 June 2002. Vol. 16, p. 321—357. DOI 10.1613/jair.953. Available from: https://jair.org/index.php/jair/article/ view/10302/24590

[32] ADEBAYO J., and KAGAL L. 2016. Iterative Orthogonal Feature Projection for Diagnosing Bias in Black-Box Models. arXiv:1611.04967 [cs, stat] [online]. 15 November 2016. [Accessed 26 April 2020]. Available from: https://arxiv.org/abs/1611.04967

[33] DING C., and PENG H. [no date][no date]. Minimum Redundancy Feature Selection from Microarray Gene Expression Data [Accessed 26 April 2020]. Available from: https://ranger.uta.edu/~chqding/papers/gene_select.pdf

[34] GINESTET C.E. [no date][no date]. Regularization: Ridge Regression and Lasso. Linear Models. Available from: http://math.bu.edu/people/cgineste/classes/ma575/p/w14_1.pdf

[35] BIAU G., and SCORNET E. 2016. A random forest guided tour. June 2016. Vol. 25, no. 2, p. 197—227. DOI 10.1007/s11749-016-0481-7. Available from: http://www.lsta.upmc.fr/BIAU/bs.pdf

[36] RYU H.J., ADAM H., and MITCHELL M. 2017. InclusiveFaceNet: Improving Face Attribute Detection with Race and Gender Diversity. arXiv:1712.00193 [cs] [online]. 1 December 2017. [Accessed 10 October 2019]. Available from: http://www.lsta.upmc.fr/BIAU/bs.pdf

[37] ZHANG B.H., LEMOINE B., and MITCHELL M. 2018. Mitigating Unwanted Biases with Adversarial Learning. In Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics and Society [online]. New Orleans LA USA: ACM. 27 December 2018. P. 335—340. [Accessed 30 September 2020]. ISBN 978-1-4503-6012-8. Available from: https://dl.acm.org/doi/10.1145/3278721.3278779

[38] ISO/IEC/IEEE 29119-1:2013 Системная и программная инженерия. Тестирование программного обеспечения. Часть 1. Понятия и определения (Software and systems engineering — Software testing — Part 1: Concepts and definitions)

[39] BEUTEL A., CHEN J., ZHAO Z., and CHI E. H. 2017. Data Decisions and Theoretical Implications when Adversarially Learning Fair Representations. arXiv:1707.00075 [cs] [online]. 30 June 2017. [Accessed 10 October 2019]. Available from: https://arxiv.org/abs/1707.00075

[40] BLOOMBERG, Amazon Doesn't Consider the Race of Its Customers. Should It? [2016]. [online]. [Accessed 30 September 2020]. Available from: https://www.bloomberg.com/graphics/2016-amazon-same-day/

[41] BONILLA-SILVA E. Racism without racists: Color-blind racism and the persistence of racial inequality in the United States. 2010. Rowman & Littlefield Publishers

[42] BONILLA-SILVA. The Structure of Racism in Color-Blind, «Post-Racial» America

[43] MITCHELL M., WU S., ZALDIVAR A., BARNES. P., VASSERMAN L., HUTCHINSON B., SPITZER E., RAJI I.D., and GEBRU T. 2019. Model Cards for Model Reporting. Proceedings of the Conference on Fairness, Accountability and Transparency — FAT '19. 2019. P. 220—229. DOI 10.1145/3287560.3287596. Available from: https://arxiv.org/pdf/1810.03993.pdf

[44] ARNOLD M., BELLAMY R.K.E., HIND M., HOUDE S., MEHTA S., MOJSILOVIC A., NAIR R., RAMAMURRTHY K.N., REIMER D., OLTEANU A., PIORKOWSKI D., TSAY J., and KUSH R.V. 2019. FactSheets: Increasing Trust in AI Services through Supplier's Declarations of Conformity. Available online: https://arxiv.org/abs/1808.07261v2

[45] Overview — ELI5 0.9.0 documentation, [no date]. [online]. [Accessed 30 September 2020]. Available from: https://eli5.readthedocs.io/en/latest/overview.html

[46] ADEBAYO J. 2020. adebayoj/fairml [online]. Python. [Accessed 30 September 2020]. Available from: https://github.com/adebayoj/fairml

[47] Using the What-If Tool | AI Platform Prediction | Google Cloud, [no date][no date]. [online]. [Accessed 30 September 2020]. Available from: https://cloud.google.com/ai-platform/ prediction/docs/using-what-if-tool

[48] RIBEIRO M.T.C. 2020. Lime: Explaining the predictions of any machine learning classifier, marcotcr/lime [online]. JavaScript. [Accessed 30 September 2020]. Available from: https:// github.com/marcotcr/lime

[49] Fairness A.I. 360, [no date]. [online]. [Accessed 30 September 2020]. Available from: http://aif360.mybluemix.net

[50] Fairlearn: A toolkit for assessing and improving fairness in AI — Microsoft Research, [no date]. [online]. [Accessed 30 September 2020]. Available from: https://www.microsoft.com/en-us/research/publication/fairlearn-a-toolkit-for-assessing-and-improving-fairness-in-ai/

[51] oracle/Skater, 2020. [online]. Python. Oracle. [Accessed 30 September 2020]. Available from: https://github.com/oracle/SkaterPythonLibraryforModelInterpretation/Explanations

[52] LUNDBERG S. 2020. slundberg/shap [online]. Jupyter Notebook. [Accessed 30 September 2020]. Available from: https://github.com/slundberg/shap A game theoretic approach to explain the output of any machine learning model

[53] Fair Test, Columbia/fairtest. Available from: https://github.com/columbia/fairtes

[54] Themis. LASER-UMASS/Themis. https://github.com/LASER-UMASS/Themis

[55] ИСО 26000 Руководство по социальной ответственности (Guidance on social responsibility)

[56] ИСО 31000 Менеджмент риска. Принципы и руководство (Risk management — Guidelines)

[57] City of Amsterdam, Algorithmic Systems of Amsterdam. [online]. [Accessed 15 July 2021]. Available from: https://algoritmeregister.amsterdam.nl/en/ai-register/

[58] City of Helsinki, Artificial Intelligence Systems of Helsinki. [online]. [Accessed 15 July 2021]. Available from: https://ai.hel.fi/en/ai-register/

УДК 006.034:006.354 ОКС 35.020

Ключевые слова: искусственный интеллект, системы искусственного интеллекта, машинное обучение, смещенность, предвзятость, принятие решений

Редактор *В.Н. Шмельков*
Технический редактор *В.Н. Прусакова*
Корректор *О.В. Лазарева*
Компьютерная верстка *Л.А. Круговой*

Сдано в набор 14.11.2023. Подписано в печать 28.11.2023. Формат 60×84⅛. Гарнитура Ариал.
Усл. печ. л. 4,65. Уч.-изд. л. 3,72.

Подготовлено на основе электронной версии, предоставленной разработчиком стандарта

Создано в единичном исполнении в ФГБУ «Институт стандартизации»
для комплектования Федерального информационного фонда стандартов,
117418 Москва, Нахимовский пр-т, д. 31, к. 2.
www.gostinfo.ru info@gostinfo.ru