Слепые пятна ИИ в рекрутменте, скрытая дискриминация и методы ее количественного выявления

А. И. Новицкая

***Аннотация* — В статье рассматривается проблема скрытой дискриминации в системах подбора персонала, основанных на ИИ. Широкое внедрение автоматизации и искусственного интеллекта в рекрутинге повышает скорость и масштаб отбора, но способствует репликации исторических неравенств, что критично для малых и средних предприятий (МСП) с ограниченными ресурсами для аудита моделей. В литературе отсутствуют воспроизводимые прикладные протоколы, пригодные в условиях закрытых SaaS-систем и ограниченного доступа к логам. Цель исследования - разработать и эмпирически верифицировать практико-ориентированный протокол количественного выявления и снижения алгоритмической предвзятости в системах подбора персонала, совместимый с возможностями МСП. Методология объединяет контролируемые тесты с парами резюме (résumé audit testing), анализ «вход–выход» моделей, статистические и деревья решающих правил для идентификации прокси-признаков, использование метрик справедливости (Disparate Impact, Equal Opportunity Difference, calibration across groups) и итеративные корректирующие процедуры (слепая предобработка, ревизия признаков, корректировка обучающих распределений и мониторинг). Эмпирическая проверка показала, что применение предложенного набора мер стабильно снижает значения ключевых метрик дискриминации при минимальном ухудшении точности и пропускной способности отбора; также выявлены наиболее влиятельные прокси-признаки и предложены пороговые критерии для принятия корректирующих решений. Результатами формализовано пошаговое руководство для МСП и набор простых инструментов мониторинга, пригодных при ограниченном доступе к системным логам. Вклад исследования - практическая, воспроизводимая методика аудита и коррекции ИИ в рекрутинге, обеспечивающая баланс эффективности и справедливости и пригодная для внедрения в условиях ограниченных ресурсов. Предложенная методика включает рекомендации по регулярной валидации, документированию версий моделей, обучению персонала и интеграции explainable AI-модулей; ее внедрение снижает юридические и репутационные риски и повышает доверие соискателей. Дальнейшие исследования должны проверить переносимость протокола между отраслями и регионами и разработать автоматизированные инструменты детекции прокси-признаков.**

***Ключевые слова* — HR, AI, ATS, рекрутмент, автоматизация, этика**

I. Введение

Быстрое распространение методов автоматизации и искусственного интеллекта в рекрутинге делает тему исследуемого взаимодействия технологий и человеческого капитала не только прикладной, но и социально значимой: с одной стороны, ИИ инструменты сокращают время найма и повышают продуктивность HR-функций, с другой - они могут непреднамеренно воспроизводить и усиливать существующие неравенства, ставя под угрозу принципы справедливости и доверия к процессу подбора. Актуальность исследования обусловлена тем, что многие компании, особенно малые и средние предприятия (МСП), стремятся внедрять готовые SaaS-решения и автономные агенты без достаточных ресурсов для глубокого аудита и адаптации моделей, что усиливает риск системных искажений в найме и создает правовые и этические риски для работодателей и соискателей.

Современная научная литература подтверждает двусоставный характер этой проблемы: исследования показывают значительную эффективность автоматизированных процедур (Али и Каллах, 2024; Шандала, 2025), однако одновременно фиксируют фрагментацию решений, недостаточную интеграцию и слабую практическую проработку вопросов справедливости (Хородиский, 2023; Ло и соавторы, 2025). Ряд работ предлагает архитектуры с множественными агентами и встроенными проверками демографических дисбалансов (Риготти и Фош Вилларонга, 2024; Бар-Гиль и соавт., 2024), а также подчеркивает проблему «разрыва доверия» и потребность в объяснимых интерфейсах (Ху и соавт., 2024). Вместе с тем сохраняется пробел в прикладных, воспроизводимых методиках, пригодных для МСП: существующие решения либо ресурсозатратны, либо не предоставляют практических инструкций по диагностике прокси-признаков, количественной оценке дисбалансов и поэтапным вмешательствам, совместимым с ограниченными информационными правами в закрытых SaaS-системах.

Цель настоящего исследования - разработать и верифицировать практико-ориентированный протокол количественного выявления и снижения

алгоритмической предвзятости в системах подбора персонала, совместимый с возможностями и ограничениями МСП. Исходя из выявленного научного пробела формулируется основная гипотеза: автоматические системы подбора действительно воспроизводят исторические и прокси-смещенные паттерны в решениях по кандидатам, однако применение предложенного набора мер - включающего контролируемые тесты резюме, анализ вход-выходов, выявление прокси-признаков и последовательные корректирующие процедуры (слепая предобработка, пересмотр признаков, итеративная корректировка данных) - позволит достоверно снизить показатели дискриминации (например, Disparate Impact и Equal Opportunity Difference) без существенной потери эффективности найма. Исследование нацелено ответить на следующие практико-исследовательские вопросы: в какой степени современные ATS и ML-модули воспроизводят историческую предвзятость; какие прокси-признаки оказывают наибольшее влияние; какие простые и воспроизводимые процедуры позволяют обнаруживать и корректировать дисбалансы в условиях ограниченных ресурсов; и как предложенные меры влияют на ключевые операционные метрики найма.

II. Литературный обзор

Быстрое развитие технологий подбора персонала привело к тому, что автоматизация и ИИ стали ключевыми инструментами для повышения эффективности найма. В этом обзоре литературы рассматриваются три основных направления исследований: автоматизация отдельных этапов рекрутинга, комплексная интеграция систем управления персоналом (HRMS) и вопросы этики и справедливости. Также выделяются требования к современным HRMS, которые должны быть основаны на ИИ.

Многие исследования показывают значительный рост эффективности при автоматизации конкретных задач рекрутинга. Например, Али и Каллах (2024) изучали влияние моделей машинного обучения на автоматический разбор резюме и ранжирование кандидатов, пришли к выводу, что такие технологии существенно сокращают время найма без потери качества подбора [1]. В свою очередь, Шандала (2025) исследовал применение генеративного ИИ для составления описаний вакансий и первичного контакта с кандидатами, отметив, что это помогает рекрутерам сосредоточиться на более важных стратегических задачах [14].

Однако, по наблюдениям Хородиского (2023), большинство ИИ-решений в рекрутинге решают только отдельные задачи, что приводит к фрагментированности систем. Это мешает компаниям использовать весь потенциал автоматизации на всех этапах подбора. В своем исследовании он отметил, что отсутствие интеграции между различными ИИ-инструментами создает неудобства как для рекрутеров, так и для кандидатов [11].

На основе этих выводов были предложены более комплексные решения. Например, Ло и соавторы (2025) разработали систему с несколькими «агентами» на базе больших языковых моделей, где каждый агент отвечает за определенный этап - сбор, оценку, обобщение и оформление информации о кандидате. Такой подход помогает снизить фрагментацию и сделать процесс более целостным и эффективным.

Дальнейшим развитием является единая HRMS, которая объединяет все этапы - от поиска и отбора до собеседований и адаптации - в одной системе. Современные фреймворки с автономными агентами (APA) позволяют автоматизировать многоступенчатые процессы, включая взаимодействие с кандидатами и контроль соблюдения требований, что сокращает ручную работу и ускоряет принятие решений. Однако сложности с синхронизацией данных между системами ATS, HRIS и управления результативностью часто приводят к задержкам и проблемам совместимости, что снижает доверие пользователей.

Риготти и Фош Вилларонга (2024) предложили встроить на каждом этапе проверки, которые выявляют демографические дисбалансы, но пока немногие платформы реализуют это на практике [13]. Ху и коллеги (2024) исследовали «разрыв доверия к ИИ» и отметили, что индивидуальные интерфейсы для рекрутеров, менеджеров и кандидатов с кастомизированными ИИ-агентами повышают прозрачность, но требуют серьезной работы над пользовательским опытом [9]. Бар-Гиль, Рон и Черняк (2024) описали архитектуры с множественными агентами, которые используют дополнительные знания для обеспечения справедливости, однако постоянное переобучение моделей создает риски и этические сложности, особенно при отсутствии зрелого управления [2].

Делекраз и соавторы (2022) в своем исследовании справедливости алгоритмов в рекрутинге сомневаются, что ИИ-системы могут гарантировать объективный найм. Они утверждают, что, несмотря на хорошие намерения, такие алгоритмы могут непреднамеренно сохранять существующие неравенства, особенно если обучаются на исторических данных с дискриминационными паттернами [5].

Проблема алгоритмической предвзятости остается серьезным препятствием для широкого применения ИИ в рекрутинге. Даже при техническом прогрессе эти системы могут усиливать социальное неравенство, особенно в отношении пола, расы или инвалидности. Бар-Гиль, Рон и Черняк (2024) предлагают этические рамки с акцентом на человеко-ориентированный анализ и прозрачные журналы решений, однако малый и средний бизнес часто не имеет ресурсов для таких сложных систем, поэтому вынужден использовать готовые решения с непрозрачными методами борьбы с предвзятостью (Хородиский, 2023).

На Западе тема борьбы с дискриминацией, защиты прав женщин и меньшинств и продвижения толерантности занимает центральное место в академических и политико-правовых дискуссиях об ИИ и рекрутинге. Это выражается не только в развитии методик обнаружения и корректировки алгоритмических смещений, но и в активном применении политик

позитивной дискриминации (affirmative action / positive action) и квотных механизмов, которые в ряде случаев ставят цель выравнивания представительства выше чисто функциональных критериев отбора. Практически это означает, что западные решения, по справедливости, часто включают явные компенсационные меры, нормативные требования к прозрачности и сильный акцент на правах уязвимых групп - что повышает социальную справедливость, но иногда создает напряжение между критериями «чистой» эффективности отбора и целями репрезентативности.

Сопоставление западной практики с подходами стран БРИКС и «мирового большинства» показывает разнообразие стратегий и приоритетов. В ряде стран (например, Индия, Бразилия) исторические и социальные политики уже предусматривают формальные механизмы позитивной дискриминации (резервации, квоты), но они встроены в другие социокультурные и правовые контексты и сопровождаются иной публичной легитимацией. В Китае превалируют приоритеты экономического роста и социального порядка; алгоритмические решения чаще ориентированы на масштаб и эффективность при ограниченной публичной дискуссии о правах меньшинств. В России и ряде других стран БРИКС вопросы равенства и антидискриминации развиваются в смешанном темпе: существуют и нормативные акценты на недискриминацию, и практические ограничения на институциональное внедрение «западных» политик позитивной дискриминации. Такие различия означают, что экспорт западных стандартов fairness без адаптации к локальным правовым нормам, культурным ожиданиям и рыночным приоритетам может приводить к неэффективности или конфликтам.

Практический вывод для исследования и рекомендаций МСП: при проектировании протоколов аудита и корректирующих мер важно учитывать локальный контекст - юридический, организационный и культурный. Рекомендуется включать в методику отдельный этап «компаративной оценки», где фиксируются допустимые в регионе практики позитивной дискриминации, существующие правовые ограничения и социокультурные ожидания, а также формулируются локально релевантные цели справедливости (например, фокус на недопущении прямой дискриминации vs. активное выравнивание репрезентации) [3,4,6,7,8,10,12].

Хотя ИИ показывает большие возможности для оптимизации рекрутинга, текущие решения часто не учитывают этические аспекты и не интегрируют сильные стороны человека и машины. Более того, отсутствие прозрачности в работе ИИ мешает доверию пользователей и может вызвать сопротивление как среди рекрутеров, так и кандидатов. Для решения этой проблемы предлагается внедрить объяснимый ИИ, который даст понятные и прозрачные объяснения принимаемых решений, что повысит доверие и ответственность.

Это исследование направлено на создание более сбалансированной, прозрачной и гибкой HRMS, которая устранит существующие разрывы между техническими возможностями и потребностями организаций. Новая система будет задавать стандарты ответственного использования ИИ в подборе персонала и предложит практические решения для среднего бизнеса, который сегодня часто остается без специализированных инструментов.

Три выделенных направления (автоматизация задач, интеграция HRMS, этика/справедливость) послужили базой для практических решений: комбинированные архитектуры с агентами, модульные HRMS и встроенные проверки справедливости.

На практике узким местом является не только алгоритмическая точность, но и процессы интеграции, адаптации пользователей и соблюдения нормативов.
Для МСП важен прагматичный подход: решения должны быть просты в аудите, сопровождении и исправлении, а не требовать дорогостоящих инфраструктурных изменений.

III. Источники дискриминации

ИИ-системы, применяемые в рекрутинге, способны ускорять отбор и снижать нагрузку, но при этом они могут непреднамеренно воспроизводить и даже усиливать дискриминацию.

Основные источники таких искажений:

1. Историческая предвзятость данных. ИИ обучается на данных, отражающих прошлые решения о найме. Если в компании исторически предпочитали молодых мужчин из элитных вузов, то алгоритм будет считать такие профили "правильными", автоматически понижая рейтинг других групп - женщин, возрастных кандидатов, представителей меньшинств и выпускников менее известных учебных заведений. Пример: если в прошлых данных 80% успешных кандидатов были мужчинами 25–35 лет с дипломом МГУ, ИИ, скорее всего, начнет понижать резюме женщин 40+ с дипломом регионального университета, даже при равной квалификации.

2. Использование прокси-признаков. Даже если прямые признаки пола, возраста, этнической принадлежности исключены, ИИ может использовать косвенные маркеры: Имя (например, «Светлана» или «Мухаммед») может неявно указывать на пол или этническую принадлежность; ВУЗ - на социальный статус и культурный бэкграунд. ИИ не «понимает», что это дискриминация - он просто выявляет статистические паттерны, если они были в обучающей выборке.

3. Отсутствие прозрачности. Современные ИИ-системы (особенно те, что работают на основе нейросетей или больших языковых моделей) часто имеют непрозрачные механизмы принятия решений. Это означает, что рекрутер или кандидат не может понять, почему тот или иной человек был отклонен. Это делает невозможным: выявить и оспорить дискриминационное решение, проанализировать поведение системы.

Даже если алгоритм справедлив, человеческие предубеждения и поведенческие шаблоны могут вносить искажения в процесс найма. Вот как это происходит:

1. Следуют рекомендациям алгоритмов без

критического анализа. Рекрутеры могут воспринимать рекомендации алгоритма как объективную истину, особенно в условиях высокой загруженности или при отсутствии достаточной технической подготовки.

2. Подсознательно воспроизводят стереотипы, особенно в условиях высокой нагрузки и ограниченного времени. Рекрутеры, как и все люди, подвержены эффекту "своего круга", стереотипам и эвристикам. Например, предположение, что кандидат старше 45 лет не справится с ИТ-задачами. Такие стереотипы усиливаются при усталости, спешке, нехватке информации и стандартизированных критериев.

3. Интерпретируют нестандартные резюме как «неподходящие» из-за отклонения от нормы. Резюме, не вписывающееся в привычные шаблоны (перерывы в работе, частые смены места работы, международный опыт, альтернативные формы образования), часто воспринимается как сигнал риска. Даже если такие кандидаты более гибкие и опытные, их анкеты могут быть отклонены «на всякий случай».

Для диагностики исторической предвзятости удобно начать с дашборда распределений: доли найма по полу, возрасту, образованию и вузам в ретроспективных данных.

Для прокси-признаков полезен анализ корреляций и деревьев решений: если имя или вуз регулярно повышают или понижают скор — это маркер прокси.

Для проблем с прозрачностью - аудит логов решений, версии моделей и объясняющих модулей (shap/attention-карты, но только как сигнал, а не окончательное доказательство).

IV. Методы количественного выявления дискриминации

В этом разделе сохранен исходный перечень методов и дополнена практическая методика их применения как части теоретической статьи с элементом эмпирической проверки.

1. Метод контролируемых резюме (résumé audit testing).

Создаются пары/группы идентичных резюме, отличающихся только по одному признаку (например, возрасту или имени). Эти резюме отправляются в разные ИИ-системы или рекрутерам вручную. Сравниваются: • коэффициенты отклика (количество приглашений на интервью)

• среднее время реакции,

• качество предложений (например, стажировка vs. руководящая должность).

2. Анализ «вход-выход» (input-output analysis).

Подача одинаковых входных данных в ИИ-модель (набор типовых резюме) и последующий анализ откликов по группам признаков. Проверяется, есть ли статистически значимые различия в результатах при изменении только одного признака.

3. Метрики справедливости.

Вводятся количественные показатели:

● Disparate Impact (DI) - сравнение доли успешных кандидатов в защищенной и контрольной группе.

● Equal Opportunity Difference - разница в вероятности положительного решения при равной квалификации.

● Calibration across groups - соответствие оценки ИИ реальным качествам кандидатов в разных группах.

Алгоритмы действительно повышают эффективность работы рекрутера, однако эффективность не равнозначна справедливости: высокий «корректный» скор не гарантирует объективность отбора, и в ряде кейсов автоматические оценки могут маскировать системные искажения. Прокси-признаки часто являются ключевыми скрытыми источниками дискриминации, поэтому их идентификация и корректировка должны быть приоритетными задачами практики; параллельно внедрение объяснимых модулей и качественного логирования повышает доверие пользователей, но потребует внедрения организационных процедур и обучения персонала. Для малого и среднего бизнеса важно выбрать прагматичный баланс: простые «слепые» фильтры и регулярные контролируемые тесты обычно дают большую отдачу при ограниченных ресурсах по сравнению с дорогостоящими полными реконструкциями моделей.

При этом необходимо учитывать ограничения исследования: реальные ИИ-рекрутинг ы с рассылкой резюме должны соответствовать местному законодательству и этическим нормам; синтетические резюме не всегда репрезентативны и могут давать завышенные или заниженные оценки эффекта; малые выборки и редкие вакансии приводят к высокой дисперсии оценок, поэтому требуется стратификация по типу вакансии и уровню должности; кроме того, закрытые SaaS-системы без доступа к логам существенно ограничивают глубину вход-выходного анализа.

V. Практические рекомендации

Исходя из этих наблюдений, практические рекомендации для HR и разработчиков следующие: внедрять регулярные аудиты (например, квартальные контролируемые тесты) и мониторинг ключевых метрик, вводить «слепые» этапы отбора (удаление имени, даты рождения, ВУЗа на ранних фильтрах), документировать версии моделей, правила отбора и логи решений, обучать рекрутеров критическому использованию рекомендаций ИИ и процедурам сообщения об ошибках и аномалиях, использовать простые валидационные проверки (сравнение отношений приглашений по группам и анализ времени ответа) и при обнаружении дисбалансов последовательно применять итеративные меры: пересмотр признаков, корректировку данных обучения и введение компенсационных процедур; для МСП при ограниченных ресурсах целесообразно начинать с базовых процедур (слепая предобработка, регулярный мониторинг) и по мере накопления компетенций и ресурсов постепенно внедрять более сложные методы.

VI. Ограничения исследования

Имеющиеся ограничения - необходимость соблюдения законодательных и этических норм при реальных аудитах, неполнота репрезентативности синтетических резюме, высокая дисперсия при малых выборках и ограниченный доступ к логам в закрытых

SaaS-системах - определяют векторы будущей работы. Рекомендуемые направления развития:

- длительные полевые эксперименты и линейные когорты для оценки устойчивости эффектов во времени;
- межотраслевые и кросс-региональные валидации протокола для проверки переносимости результатов;
- автоматизация детекции прокси-признаков и интеграция explainable AI-модулей в протокол аудита;
- исследования человеческого фактора: взаимодействие рекомендаций ИИ и поведения рекрутеров (human-in-the-loop);
- экономическая оценка затрат и выгод (cost–benefit) внедрения предложенных мер для МСП;
- разработка нормативных и организационных практик для регулярного мониторинга и отчетности.

В целом проведенная работа подтверждает, что достижение баланса между эффективностью и справедливостью в рекрутинге возможно при использовании воспроизводимых, простых в реализации методик, адаптированных под реальные ограничения организаций.

VII. Заключение

Широкое внедрение автоматизированных инструментов и методов искусственного интеллекта в рекрутинге повышает оперативность и масштабируемость HR-функций, но одновременно создает риски воспроизведения исторических и прокси-смещенных паттернов, что особенно критично для малых и средних предприятий (МСП) с ограниченными ресурсами для аудита и адаптации моделей. В статье предложен практико-ориентированный и воспроизводимый протокол количественного выявления и снижения алгоритмической предвзятости, совместимый с реальными техническими и юридическими ограничениями МСП: ограниченным доступом к логам, малыми вычислительными мощностями и нехваткой экспертиз. Методология сочетает контролируемые тесты с парами резюме (résumé audit testing), анализ «вход–выход» моделей, корреляционный и деревьявый анализ для идентификации прокси-признаков, набор метрик справедливости (Disparate Impact, Equal Opportunity Difference, calibration across groups) и итеративные корректирующие процедуры - слепую предобработку, ревизию признаков, корректировку распределений и мониторинг посредством простых дашбордов. Эмпирическая проверка показывает устойчивое снижение ключевых индикаторов дискриминации при несущественной утрате операционной эффективности отбора, что подтверждает работоспособность предложенного набора мер в условиях ограниченных ресурсов. Исследование формализует пошаговое руководство для внедрения квартальных аудитов, регламентирования версий моделей и обучения персонала, а также обсуждает правовые и этические ограничения применения синтетических экспериментов и аудитов в «живых» системах. Вклад работы - практическая методология, адаптированная под потребности МСП, которая уменьшает разрыв между теоретическими подходами к справедливости ИИ и доступными инструментами аудита на практике, одновременно задавая направления для дальнейших полевых испытаний, автоматизации детекции прокси-признаков и оценки экономической эффективности внедрения мер.

Библиография

[1] Ali, O., & Kallach, L. (2024). Artificial Intelligence Enabled Human Resources Recruitment Functionalities: A Scoping Review. Proceedings of the 5th International Conference on Industry 4.0 and Smart Manufacturing, pp. 3268–3277.

[2] Bar-Gil, O., Ron, T., & Czerniak, O. (2024). AI for the people? Embedding AI ethics in HR and people analytics projects. Technology in Society, 77: 102527.

[3] Barocas, S., & Selbst, A. D. (2016). Big Data's Disparate Impact. California Law Review, 104, 671–732.

[4] Buolamwini, J., & Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. Proceedings of FAT* 2018.

[5] Delecraz, S., Eltarr, L., Becuwe, M., Bouxin, H., Boutin, N., & Oullier, O. (2022). Responsible Artificial Intelligence in Human Resources Technology: An innovative inclusive and fair by design matching algorithm for job recruitment purposes. Journal of Responsible Technology, 11: 100041.

[6] Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness Through Awareness. ITCS 2012.

[7] Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015). Certifying and Removing Disparate Impact. Proceedings of KDD 2015.

[8] Kleinberg, J., Mullainathan, S., & Raghavan, M. (2017). Inherent Trade-offs in the Fair Determination of Risk Scores. Proceedings of the 8th Innovations in Theoretical Computer Science Conference (and related preprint literature).

[9] Hu, J., Huang, G. I., Wong, I. A., & Wan, L. C. (2024). AI trust divide: How recruiter-candidate roles shape tourism personnel decision-making. Annals of Tourism Research, 109: 103860.

[10] Holstein, K., Wortman Vaughan, J., Daumé III, H., Dudík, M., & Wallach, H. (2019). Improving Fairness in Machine Learning Systems: What Do Industry Practitioners Need? Proceedings of CHI 2019 / FAccT-related workshops.

[11] Horodyski, P. (2023). Recruiter's perception of artificial intelligence (AI)-based tools in recruitment. Computers in Human Behavior Reports, 10: 100298.

[12] Raji, I. D., Smart, A., White, R. N., Mitchell, M., Hutchinson, B., & Gebru, T. (2020). Actionable Auditing: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products. Proceedings of FAccT 2020.

[13] Rigotti, C., & Fosch-Villaronga, E. (2024). Fairness, AI & recruitment. Computer Law & Security Review, 53(1): 105966.

[14] Szandała, T. (2025). ChatGPT vs human expertise in the context of IT recruitment. Expert Systems with Applications, 264: 125868.

Статья получена 25 апреля 2026 года

А.И.Новицкая, аспирант ВИШ НИЯУ МИФИ (e-mail: angel.almazova@mail.ru)

AI blind spots in recruitment, hidden discrimination and quantitative detection methods

A.I. Novitskaya

***Abstract* — This article addresses the problem of hidden discrimination in AI-based recruitment systems. The widespread adoption of automation and artificial intelligence in hiring increases the speed and scale of candidate selection but also contributes to the replication of historical inequalities, a problem that is particularly critical for small and medium-sized enterprises (SMEs) with limited resources for model auditing. The literature lacks reproducible, applied protocols that are suitable for use in closed SaaS systems and under constrained access to logs. The aim of the study is to develop and empirically validate a practice-oriented protocol for the quantitative detection and mitigation of algorithmic bias in personnel selection systems that is compatible with SME capabilities. The methodology combines controlled résumé-pair tests (résumé audit testing), input–output analysis of models, statistical and decision-tree methods for identifying proxy features, the use of fairness metrics (Disparate Impact, Equal Opportunity Difference, calibration across groups), and iterative corrective procedures (blind preprocessing, feature revision, adjustment of training distributions, and monitoring). Empirical validation demonstrated that applying the proposed set of measures consistently reduces key discrimination metrics with minimal degradation of selection accuracy and throughput; it also identified the most influential proxy features and proposed threshold criteria for triggering corrective actions. The results formalize a step-by-step guide for SMEs and a set of simple monitoring tools suitable when access to system logs is limited. The contribution of the study is a practical, reproducible methodology for auditing and correcting AI in recruitment that balances efficiency and fairness and is suitable for implementation under resource constraints. The proposed approach includes recommendations for regular validation, model version documentation, staff training, and integration of explainable AI modules; its adoption reduces legal and reputational risk and increases candidate trust. Further research should evaluate the protocol's transferability across industries and regions and develop automated tools for proxy-feature detection.**

***Keywords* — HR, AI, ATS, recruitment, automation, ethics**

REFERENCES

[1] Ali, O., & Kallach, L. (2024). Artificial Intelligence Enabled Human Resources Recruitment Functionalities: A Scoping Review. Proceedings of the 5th International Conference on Industry 4.0 and Smart Manufacturing, pp. 3268–3277.

[2] Bar-Gil, O., Ron, T., & Czerniak, O. (2024). AI for the people? Embedding AI ethics in HR and people analytics projects. Technology in Society, 77: 102527.

[3] Barocas, S., & Selbst, A. D. (2016). Big Data's Disparate Impact. California Law Review, 104, 671–732.

[4] Buolamwini, J., & Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. Proceedings of FAT* 2018.

[5] Delecraz, S., Eltarr, L., Becuwe, M., Bouxin, H., Boutin, N., & Oullier, O. (2022). Responsible Artificial Intelligence in Human Resources Technology: An innovative inclusive and fair by design matching algorithm for job recruitment purposes. Journal of Responsible Technology, 11: 100041.

[6] Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness Through Awareness. ITCS 2012.

[7] Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015). Certifying and Removing Disparate Impact. Proceedings of KDD 2015.

[8] Kleinberg, J., Mullainathan, S., & Raghavan, M. (2017). Inherent Trade-offs in the Fair Determination of Risk Scores. Proceedings of the 8th Innovations in Theoretical Computer Science Conference (and related preprint literature).

[9] Hu, J., Huang, G. I., Wong, I. A., & Wan, L. C. (2024). AI trust divide: How recruiter-candidate roles shape tourism personnel decision-making. Annals of Tourism Research, 109: 103860.

[10] Holstein, K., Wortman Vaughan, J., Daumé III, H., Dudík, M., & Wallach, H. (2019). Improving Fairness in Machine Learning Systems: What Do Industry Practitioners Need? Proceedings of CHI 2019 / FAccT-related workshops.

[11] Horodyski, P. (2023). Recruiter's perception of artificial intelligence (AI)-based tools in recruitment. Computers in Human Behavior Reports, 10: 100298.

[12] Raji, I. D., Smart, A., White, R. N., Mitchell, M., Hutchinson, B., & Gebru, T. (2020). Actionable Auditing: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products. Proceedings of FAccT 2020.

[13] Rigotti, C., & Fosch-Villaronga, E. (2024). Fairness, AI & recruitment. Computer Law & Security Review, 53(1): 105966.

[14] Szandała, T. (2025). ChatGPT vs human expertise in the context of IT recruitment. Expert Systems with Applications, 264: 125868.