Гюнай Казимзаде
Аспирантка в Институте сетевого общества им. Дж.Вейценбаума (Германия)

Технологии разнообразия против технологий дискриминации на примере систем, основанных на искусственном интеллекте

Вступление

Сегодня системы искусственного интеллекта используются в различных областях и, если посмотреть совсем свысока, эти технологии функционируют как *системы дискриминации*. Они дифференцируют, ранжируют, категоризируют, и, таким образом, в некоторых особых случаях, они дискриминируют и создают общественное неравенство.

По мере того, как современные системы распознавания лиц неправильно категоризируют людей разных цветов кожи, женщин рекомендуют на менее оплачиваемые позиции, а автоматические рекрутинговые системы исключают кандидаток из конкурса на технические и руководящие позиции, общество стоит перед лицом вызова «категоризации», «дискриминации» и «несправедливой оценки» алгоритмическими системами[171, 172, 173, 174].

Авторы доклада AI Now Institute отмечают наличие кризиса *разнообразия*[175] в технологиях искусственного интеллекта (здесь и далее — ИИ) в таких областях, как гендер и раса[176]. Авторы ведущих конференций по ИИ, руководители, сотрудники и исследовательский персонал «техно гигантов» Google, Facebook, Microsoft — преимущественно белые мужчины. Авторы доклада отмечают отсутствие общественно доступных данных о трансгендерных сотрудниках и представителях других гендерных меньшинств.

171 *O'Neil C.* (2017) Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy (01 edition ed.). London: Penguin.

172 *Rice L., Swesnik D.* (2013) Discriminatory Effects of Credit Scoring on Communities of Color // 45 Suffolk University Law Review. №935. C.32.

173 *Whittaker M., Crawford K., Dobbe R., Genevieve F., Kaziunas E., Varoon M., West S.M., Richardson R., Schultz J., Schwartz O.* (2018) AI Now Report 2018 // AI Now Institute, New York University, 2018.

174 *Buolamwini J., Gebru T.* (2018) Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification // Conference on Fairness, Accountability, and Transparency.

175 Разнообразие (diversity) — широкое социологическое и моральное понятие, предполагающее уважение и принятие индивидуальных и групповых различий, таких как раса, этническая принадлежность, сексуальная ориентация, возраст, политические и религиозные взгляды и др. — *прим.ред.*

176 *West S.M., Whittaker M., Crawford K.* (2019) Discriminating Systems: Gender, Race and Power in AI // AI Now Institute. URL: https://ainowinstitute.org/discriminatingsystems.pdf (дата обращения 25.08.2020).

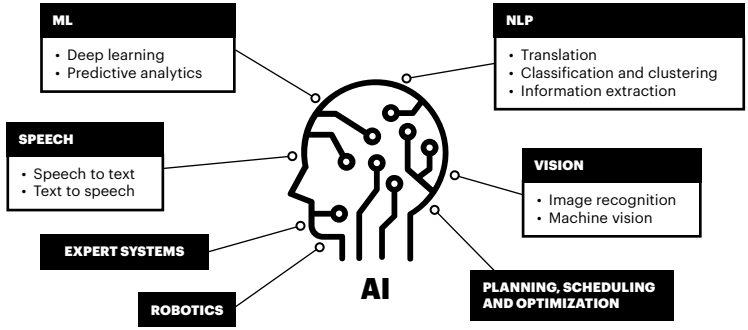

Рисунок 9. Какие технологии, основанные на ИИ, испытывают кризис разнообразия?[178]

Несмотря на то, что озабоченность и общественное внимание на «решении» проблемы разнообразия в ИИ через работу с качеством данных, справедливыми моделями и инклюзивным дизайном, многие специалисты настаивают, что необходим более глубокий анализ культуры поведения на рабочем месте, неравномерностей во властных полномочиях, домогательств, дискриминационных практик найма персонала и несправедливой оплаты труда, которые заставляют людей покидать или вообще избегать работу в секторе ИИ[177].

Представляется, что проблема неравенства в ИИ — не только техническая. Её необходимо решить используя междисциплинарный подход, задействуя различные заинтересованные стороны, политиков и руководителей, но самое главное — гражданское общество.

Технологии, основанные на ИИ, быстро позиционируются в центре человеческой жизни развивая новые горизонты для общества. Этот модный термин «ИИ» используется для обобщения технологий и систем, которые имитируют человеческий разум, используя набор таких техник, как автоматическое распознавание речи, распознавание изображений, естественную обработку языка, генерирование речи и т. д. *Машинное обучение* — часть технологий, внутри технологий искусственного интел-

177 *West S. M., Whittaker M., Crawford K.* (2019) Discriminating Systems: Gender, Race and Power in AI.

лекта, сконцентрированная на возможности машин использовать большие объемы данных для собственного обучения без необходимости дополнительного программирования со стороны человека.

Экспертные системы, также включенные в общий зонтичный термин ИИ, могут работать с общими техниками программирования с или без использования алгоритмов машинного обучения. Важно различать между собой искусственный интеллект, машинное обучение, и другие термины, роль которых разнится по своему объему и влиянию на проблему разнообразия и дискриминации в ИИ. В нашей статье мы рассматриваем все перечисленные технологии и дискриминационные проблемы, вызванные этими технологиями.

Дискуссия на предмет того, улучшает ли ИИ нашу жизнь или увеличивает неравенство, остра как никогда. Широкомасштабное машинное обучение и технологии глубокого обучения, которые позволяют компьютерам обрабатывать и анализировать большие объемы данных широко используются в таких значимых областях, как страхование (особенно, кредитный скоринг), банковские займы, здравоохранение, включая аналитику здравоохранения, роботизированное здравоохранение, постановка медицинского диагноза, общественная безопасность, особенно в области предиктивного контроля преступности, в области замены человеческого труда, управления кадрами, так же как и в областях социальных сетей, игр, сервисов цифровых развлечений, области образования для роботов педагогов, взаимодействий ребенок-робот, умных репетиторских систем, онлайн обучения и образовательной аналитики[179, 180, 181, 182].

178 Источник: Deloitte Insights. URL: deloitte.com/insights.

179 *O'Neil C.* (2017) Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy.

180 *Rice L., Swesnik D.* (2013) Discriminatory Effects of Credit Scoring on Communities of Color.

181 *Buolamwini J., Gebru T.* (2018) Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification.

182 *Whittaker M., Crawford K., Dobbe R., Genevieve F., Kaziunas E., Varoon M., West S. M., Richardson R., Schultz J., Schwartz O.* (2018) AI Now Report 2018.

В особенных случаях, эти социо-технологические системы приносят несправедливые, неэтичные и дискриминирующие результаты. Доклад AI Now Institute утверждает, что «Системы, использующие физическую внешность как показатель черт характера или внутреннего состояния глубоко неблагонадежны. Это же касается ИИ инструментов, заявляющих о способности определить сексуальность по фотографии головы, предсказывать «криминальность» по чертам лица или оценивать компетентность сотрудника по «микровыражениям». Такие системы реплицируют паттерны расовых и гендерных предрассудков, возможно углубляя и обосновывая историческое неравенство. Коммерческое применение таких инструментов является предметом сильного беспокойства»[183].

Скандал, связанный с «сексистским» рекрутинговым ИИ инструментом от Amazon, который «научился» исключать кандидатов женского пола при найме на работу, привлек внимание как общественности, так и самой компании в 2018. Причиной такого несправедливого суждения системы стали исторические данные о решениях, принятых рекрутерами за последние 10 лет. В течение этого периода очень мало кандидатов женского пола было нанято на руководящие или технические позиции; таким образом, система, натренированная на этих данных, научилась имитировать предвзятые решения, сделанные сотрудниками. После того, как новости о скандале стали приобрели характер вируса, компания решила отредактировать программу так, чтобы нейтрализовать влияние гендерных черт, хотя это и не гарантия того, что рекрутинговая система не будет коррелировать другие черты с гендерными атрибутами кандидатов.

Тимнит Гебру, исследователь в области этики алгоритмов в Microsoft, подчеркивает[184], что глубокое обуче-

183 *Olteanu A., Castillo C., Diaz F., Kiciman E.* (2016) Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries // SSRN Electronic Journal. Frontiers in Big Data. №2:13. 2016. 20 декабря. URL: http://dx.doi.org/10.2139/ssrn.2886526 (дата обращения 25.08.2020).

184 *Gebru T., Morgenstern J., Vecchione B., Vaughan J.W., Wallach H., Hal Daumeé III, Crawford K.* (2018) Datasheets for Datasets // arXiv.org [Элек-

ние может изменить рынок страховых услуг, на котором меньшинства и другие уязвимые группы могут подвергнуться дискриминации только из-за того, что автомобильные аварии происходят в густонаселенных районах, там где эти группы, в основном, проживают. Программа глубокого обучения может «выучить», что существовала значимая корреляция между принадлежностью к определенному меньшинству (например, по цвету кожи) и большим количеством автомобильных аварий. На основе этих данных, будет построена модель с заранее встроенными предрассудками о повышенной опасности людей определенного цвета кожи. Таким образом, подобная страховая система развила бы расовые предрассудки, основываясь на корреляции, а не на причинно-следственных связях.

Алгоритмы машинного обучения в системах, основанных на ИИ, применяются в здравоохранении для анализа больших объемов данных, в целях улучшения принятия решений и увеличения эффективности. Данные, собранные за годы измерений, могут отражать исторические предрассудки против уязвимых групп. Это может привести к дальнейшему воспроизводству предрассудков, что соответственно ведет к усилению неравенства в обеспечении здравоохранением.

Предиктивные полицейские алгоритмы становятся очень популярными среди городов в США, также как и в других странах. Многие ученые и исследователи частной жизни озабочены критическими последствиями решений, принятых ИИ-системами, которые потенциально способны усилить расовые и культурные предрассудки. «Полиция в США систематически предубеждена против цветных сообществ,» — заявил Fast Company юридический директор Профсоюза гражданских свобод Нью-Йорка Кристофер Данн. «Любая предиктивная полицейская платформа имеет риск наличия унаследованных диспропорций из-за чрезмерного наблюдения полиции за цвет-

тронный ресурс]. 2018. 9 июля. URL: http://arxiv.org/abs/1803.09010 (дата обращения 25.08.2020).

ными сообществами, которые послужили исходными данными для этой платформы. Чтобы обеспечить справедливость, полиция Нью-Йорка должна быть прозрачна говоря о внедряемых технологиях и позволять независимым исследователям проверять эти системы перед тем, как они будут тестироваться на жителях Нью-Йорка».[185]

Одновременно со взрывом умных технологий, социальные платформы, изначально призванные связывать людей друг с другом, стали доверенными местами для обмена личной информацией, фотографиями, досугом; в них обсуждаются темы, связанные с политикой, религией, так и другими деликатными темами. Чтобы масштабировать свою деятельность, эти платформы применяют искусственный интеллект для фильтрации сообщений (т. н. «умная лента») и таргетированной рекламы, рекомендуют кино, музыку, новостных каналы. Эти рекомендации формируются с учетом демографической информации о пользователе, поле, возрасте и истории посещения сторонних веб-страниц, таким образом предоставляя пользователям информацию, которая совместима с их уже существующим «инфо-пузырем».

Со временем, предрассудки и предубеждения этих «инфо-пузырей» усиливаются и распространяются далее. Эта же проблема встречается в голосовых интерфейсах, таких как Amazon Alexa, Microsoft Cortana или др. Растущее число пользователей испытывают на себе посредничество подобных систем, управляемых умными алгоритмами. Существует опасность того, что эти технологии будут ограничивать наши решения и взаимодействия, даже без нашего ведома.

Принимая во внимание скорость, с которой тестируются и оцениваются процессы алгоритмического принятия решений, велика вероятность, что в скором времени они будут влиять на общество еще сильнее. Жизненно необходимо, чтобы гражданское общество открыто обсуждало такие проблемы, как предрассудки и дискрими-

185 *O'Neil C.* (2017) Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy.

нация в ИИ-системах, вместе с обсуждением стратегии, видения, и плана действий по преодолению этих проблем.

Возможные направления развития и желаемое будущее

Но как? Какое возможное развитие может встать перед лицом общества при экспоненциальном росте объема данных и создания различного рода систем, основанных на ИИ, в особенности, социо-технологических систем?

Первый шаг к решению дискриминационных проблем в ИИ возможен благодаря применению рекомендаций, обращающих внимание на гендерные и культурные различия, по справедливому сбору и обработке данных, дизайну и имплементации слоев в процессе создания и функционирования ИИ-систем. Более того, жизненно важно, чтобы все общественные акторы, включая государственные организации, коммерческие компании, НКО, образовательные учреждения, следовали этим руководствам и использовали их в своих областях деятельности.

Другое направление решения проблемы неравенства состоит в применении новых технологий самим гражданским обществом. На данный момент, приложения, зависимые от данных и натренированные на неполных наборах данных, полученных для ограниченных культурных или географических групп, могут иметь предвзятые результаты по отношению к группам, не попавшим в эти наборы данных. Такое происходит из-за недоступности качественных данных в определенных территориях. Например, в базе данных, представленной UK Biobank, которая была нацелена на генотипирование 500 000 человек, недопредставлены этнические меньшинства, включая черных (на треть), китайцев (более, чем на треть), индийцев, пакистанцев (более, чем на половину). Белые британские участники составляют 94,6% выборки Biobank, по сравнению с 91,3% общего населения. Эта выборка значительно влияет на медицинские диагнозы

и создает предубеждения и неверные диагнозы среди указанных недопредставленых групп.[186] Разработчики ИИ ориентируют дизайн систем на европейское и американское население, из-за нехватки качественных данных для репрезентации населения других стран[187]. Инициативы открытых данных могут быть уникальной возможностью для включения недопредставленых групп в повестку технологических решений для повышения разнообразия.

Применение машинного обучения и ИИ технологий в области образования, а особенно роль развивающихся технологий в уменьшении социального неравенства, практически не обсуждаются. Но все же, у них есть обнадеживающий потенциал в преодолении проблемы общественного неравенства, вызванного развивающимися технологиями.

При текущем быстром темпе развития технологий, решения относительно разработки технологий и сбора данных принимаются небольшой элитой. Следовательно, существует возможность распределить знание и силу среди всех слоев общества. Это возможно через обучение нового поколения женщин-технологических лидеров, преподавание новых технологий в раннем возрасте, преподавание междисциплинарной, межкультурной кооперации, разнообразия, равно как и просвещение в области сознательных и несознательных предубеждений, внедряемых в технологии, которые, в свою очередь, влияют на общество.

Гражданское общество должно играть в этом процессе ключевую роль, понимая и адаптируя технологии, использующие данные, и их влияние на тайну частной жизни, политику и экономику, а также возможности и риски, которые они несут.

Что же касается таких проблем, как профилирование социальных медиа, политической манипуляции, дискри-

186 *Swanson J. M.* (2012) The UK Biobank and selection bias // The Lancet [Электронный ресурс]. 2012. 14 июля. URL: https://doi.org/10.1016/S0140–6736 (12) 61179–9 (дата обращения 25.08.2020).

187 *Olteanu A., Castillo C., Diaz F., Kiciman E.* (2016) Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries.

минирующих систем по принятию решений, то гражданское общество могло бы послужить мостом между обществом и законодателями, техническими исполнителями, в привлечении к диалогу представителей тех сообществ, которым они служат, включая недопредставленные уязвимые группы.

Самая уместная роль гражданского общества состоит в понимании опасностей предубеждений, вызванных системами ИИ и тем, как они могут усугубить проблемы, решить которые они призваны, при понимании возможностей этих технологий послужить обществу. Целью некоммерческих организаций может быть информирование и просвещение компаний и организаций, разрабатывающие новые алгоритмы, а также законодателей, принимающих новые законы об управлении этими технологиями, о том, что такое равенство, прозрачность, подотчетность. Просвещение может идти вместе с мониторингом и анализом последствий применяемых стратегий и стандартов.

Что может пойти не так?

Без надлежащих ограничений, системы, основанные на ИИ, могут принести негативные общественные последствия за счет создания систем централизованного манипулирования, фильтрации и дискриминации уязвимых общественных групп. Манипулируя данными для обучения и дизайном самих моделей машинного обучения в таких важных областях, как трудоустройство, подобные системы могут создавать неравенство доступа, в том числе, по географическому признаку. Это может привести к использованию таких технологий для неравномерного распределения власти и создания «разъединенных» пузырей в обществе[188]. Технологии ИИ также могут принести проблемы, связанные с тайной личной жизни, фейковых новостей и политического манипулирования через

188 *O'Neil C.* (2017) Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy.

таргетированную рекламу и фильтрацию в социальных медиа — все это является рисками подхода «разделяй и властвуй», угрожающими демократии.[189]

Децентрализация или централизация власти в этих технологиях может принести последствия, касающиеся общественной справедливости, равенства, и, следовательно, может быть использована как инструмент для манипуляции, управления и направления дальнейшего развития общества. Китайская система социального кредита сравнивается с сериалами «Черное зеркало», «Большой брат» и другими научно-фантастическими сюжетами о мрачном будущем. «Что по-настоящему страшно, так это то, что ты ничего не можешь с этим поделать. Ты не можешь никому пожаловаться. Ты застреваешь посередине пустоты» говорит один из журналистов из Китая, внесенный в «черный список» и «помеченный» как «негодный» для покупки билета на самолет, а также путешествия некоторыми железнодорожными линиями, покупки недвижимости или получения кредита[190].

Неизвестные

Обсуждения на темы прозрачности, справедливости и подотчетности алгоритмов и технологий, использующихся для влияния на жизни общества, не утихают. Но, до сих пор, результаты этих обсуждений не внедрены в дизайн и воплощение этих технологий. Не все алгоритмы машинного и глубоко обучения могут объяснить свои решения, большая часть данных, использованных для разработки таких систем, предвзяты, и не отражают все многообразие населения. Как будут управляться алгоритмы, созданные по модели «черных ящиков», как стереотипы ИИ повлияют

189 *Bourdieu P.* (1989) Social Space and Symbolic Power // American Sociological Association: Sociological Theory. №7 (1). С.14–25. URL: https://doi.org/10.2307/202060 (дата обращения 25.08.2020).

190 *Matsakis L.* (2019) How the West Got China's Social Credit System Wrong // Wired [Электронный ресурс]. 2019. 29 июля. URL: https://www.wired.com/story/china-social-credit-score-system/ (дата обращения 25.08.2020).

на тех людей, которым они призваны помочь, кто должен отвечать за управление всеми этими технологиями — на все эти вопросы ответов пока еще нет. И всё же, продолжаются инициативы институтов AI Now institute, Alan Turing Institute, Leverhulme Institute for the Future of Intelligence и других, одновременно привлекающих внимание как к критическим проблемам дискриминации и исключения в системах, использующих данные, так и к применению новых технологических решений, направленных на устранение влияния проблем, привнесенных разными законодателями и комиссиями в стратегии и политике ИИ.

Слабые сигналы

Хоть мы и спорили о стереотипах и дискриминации, создаваемых ИИ, возможно мы сможем использовать эти же технологии для выявления всех тех предубеждений, о которых мы говорили. Это одно из направлений, которое имеет медленное, но постоянное влияние на преодоление проблемы ИИ-неравенства.

Например, команда разработчиков из США создала инструмент ИИ для определения стереотипов и предубеждений, основанных на расе или гендере при принятии человека в университет или на работу. Система натренирована на большом объеме данных и дает рекомендации по найму кандидатов-женщин, если они долгое время были недопредставлены на определённых позициях или кафедрах в течение долгого времени.

Такие типы техник могут управляться и использоваться организациями гражданского общества, для того, чтобы измерять и устранять стереотипы и дискриминацию в социотехнических системах.