А. И. СМИРНОВА

ПРЕДВЗЯТОСТЬ КАК ПРОБЛЕМА АЛГОРИТМОВ ИИ: ЭТИЧЕСКИЕ АСПЕКТЫ*

В статье рассматриваются причины, способные вносить предвзятость в функционирование алгоритмов искусственного интеллекта (ИИ). Показано, что нормативного регулирования этичности недостаточно для преодоления проблемы; необходимо выявлять и корректировать случаи «усвоения» алгоритмами ИИ предвзятости во время обучения на данных, включающих предвзятые решения, принимавшиеся в реальной жизни.

Ключевые слова: *искусственный интеллект, предвзятость, этическое регулирование, обучение, коррекция.*

The article discusses the reasons that can introduce bias into the functioning of AI algorithms. It is shown that normative regulation of ethics is not enough to overcome the problem; it is necessary to identify and correct cases when AI algorithms "learn" bias during training on data that includes biased decisions made in real life.

Keywords: *artificial intelligence, bias, ethical regulation, learning, correction.*

Введение

Искусственный интеллект (ИИ) – это общее, «зонтичное» понятие, охватывающее значительное число явлений и имеющее взаимовлияние со множеством аспектов человеческой жизни. Настоящее время можно считать принципиально значимым в развитии ИИ, поскольку от выполнения конкретных технических задач с изна-

* ***Для цитирования:*** Смирнова А. И. Предвзятость как проблема алгоритмов ИИ: этические аспекты // Философия и общество. 2023. № 3. С. 118–126. DOI: 10.30884/jfio/2023.03.07.

For citation: Smirnova A. I. Bias as a Problem with AI Algorithms: Ethical Aspects // Filosofiya i obshchestvo = Philosophy and Society. 2023. No. 3. Pp. 118–126. DOI: 10.30884/jfio/2023.03.07 (in Russian).

чально запрограммированным набором вариантов развития событий и предписаний конкретных действий при каждом из них ИИ переходит к решению сложных когнитивных задач в социальной сфере с множеством возможных последствий, выбор и принятие решений в которых ранее были доступны лишь человеческому разуму.

Этот процесс, в котором можно увидеть уподобление ИИ человеческому интеллекту, вызывает бурную дискуссию в научном сообществе – и одними из наиболее спорных остаются этические вопросы, поскольку существуют опасения, что такой всеобщий ИИ приведет к сверхразуму и значительно превзойдет когнитивные способности человека практически во всех областях жизни [Bostrom 2014: 22]. Опасения, касающиеся возникновения сверхразума, строятся на идее того, что если бы мы, люди, могли создать ИИ с интеллектуальными способностями примерно на уровне человека, то это творение, в свою очередь, могло бы создать еще более высокий интеллект, который в свою очередь мог бы создать еще более высокий интеллект, и т. д. [Müller, Bostrom 2016]. Суть же таких опасений сводится к тому, что, чем более совершенным будет ИИ, тем больше рисков (вплоть до экзистенциальных) он принесет человечеству [Hawking *et al.* 2014]. Например, ИИ может вызвать массовую безработицу, принимать решения, которые люди не способны понять и контролировать, привести к перераспределению богатства и в конечном итоге заменить людей [Siau, Wang 2018]. Однако существуют и «ИИ-оптимисты», в частности Р. Курцвейл, предсказывающий наступление сингулярности в 2045 г.

Предвзятость алгоритмов ИИ: причины и последствия

Некоторые вызовы этике ИИ очень похожи на проблемы, связанные с проектированием различных предметов и устройств. Разработка робота, собирающего трубы на буровой установке, не более сложна с моральной точки зрения, чем разработка офисной мебели. Однако когда алгоритмы ИИ берут на себя когнитивную работу с социальными аспектами – когнитивные задачи, ранее выполнявшиеся людьми, – возникает острая необходимость соблюдать социальные требования. Действительно, «некритичный импорт логических алгоритмов оптимизации процесса управления и принятия решений из сферы техники и инженерии в некоторые области со-

циальных наук порождает алгоритмическую предвзятость в значительных масштабах» [Харитонова и др. 2021: 495].

Развитие систем на основе ИИ поднимает этические вопросы нового типа, касающиеся, в частности, «их влияния на процессы принятия решений, проблему занятости и рынок труда, взаимодействие между людьми в обществе, медицину, образование, средства информации, доступ к информации, цифровое неравенство, защиту персональных данных и потребителей, окружающую среду, демократию, верховенство закона, обеспечение безопасности и правопорядка, двойное использование, а также права человека и основные свободы, включая свободу выражения мнений, неприкосновенность частной жизни и отсутствие дискриминации» [ЮНЕСКО 2021].

Одним из основополагающих аргументов в пользу развития ИИ в целом и автоматизации принятия решений в частности является предполагаемое снижение предвзятости, присущей человеку и якобы не присущей алгоритмам [Bozdag 2013]. Однако к настоящему времени накоплен значительный объем научной литературы, убедительно доказывающей несостоятельность этого тезиса.

Отметим в первую очередь, что на дизайн и функциональность алгоритма оказывают фундаментальное влияние ценности его разработчиков и той культуры, к которой они принадлежат, – так, в гипотетическом, но очень ярком и образном примере, приводимом Н. Бостромом и Э. Юдковским, отмечается, что если бы алгоритмы ИИ разрабатывал Архимед, то они бы несли отпечаток древнегреческого социального устройства и культурных и этических норм, в том числе представления о том, что рабство – это абсолютно приемлемое явление (поскольку таковы были древнегреческие этические нормы) [Bostrom, Yudkowsky 2018; Железнов 2021]. Таким образом, «ценности автора [алгоритма] вольно или невольно замораживаются в коде, фактически институционализируя эти ценности» [Macnish 2012: 158]. При этом получение «очищенных» от человеческого влияния данных для тренировки алгоритмов ИИ является отдельной весьма сложной задачей, а «предоставление реальных необработанных данных по поведению людей для обучения ИИ может привести к усилению структурной дискриминации путем воспроизведения стереотипов и аттитюдов, заложенных в исходных данных» [Шиллер 2020: 100].

Более того, в широкой перспективе процессы развития – как на уровне обществ, так и на уровне отдельных индивидов – являются нелинейными, и в каждый момент времени следование этих процессов тем или иным путем не является инвариантным – напротив, имеется множество вариантов развития, которые могут субъективно ощущаться более правильными или менее правильными, лучшими или худшими, – во многом в зависимости от этических норм их наблюдателей и/или участников. Иными словами, не существует единственно правильного выбора, который должен быть по умолчанию заложен в алгоритм. Выбор того или иного варианта из множества возможных, таким образом, зачастую субъективен, а значит, может приводить к возникновению предвзятости в функционировании алгоритма. При этом, если алгоритм построен, к примеру, на основе нейросетей глубокого обучения, точные причины возникновения предвзятости в алгоритме и конкретный шаг, на котором она возникает, может быть практически невозможно отследить и выявить, – это будет неочевидно даже для самих разработчиков алгоритма.

В системах на основе ИИ могут использоваться различные методологии, в частности алгоритмы машинного обучения, в том числе глубокое обучение и обучение с подкреплением. Алгоритмическое принятие решений и интеллектуальный анализ данных зачастую базируются на корреляциях, выявленных в тех или иных наборах данных, – и трудность здесь заключается в том, что корреляция, в отличие от функциональной зависимости, необязательно заключает в себе причинно-следственную связь. Существуют (и весьма распространены) ложные корреляции, не заключающие в себе никаких подлинных причинно-следственных связей. Более того, корреляции, установленные в больших наборах данных, часто невозможно воспроизвести или опровергнуть, что еще более затрудняет поиск причинно-следственных связей. Даже если обнаружены сильные корреляции, они могут наблюдаться только на уровне населения в целом или каких-то значительных его групп, тогда как действия, решение о которых принимается на основе данных наблюдений, направлены на отдельных лиц (отметим, что это отчасти касается и функциональных причинно-следственных зависимостей – если они наблюдаются на популяционном уровне, это не означает, что их можно в неизменном виде проецировать и на микроуровень отдельных индивидов). Несмотря на это, корреляции, основанные на достаточном объеме данных, все чаще рассматриваются как впол-

не достоверные и приемлемые для того, чтобы служить основой для принятия решений и выполнения соответствующих действий без предварительного установления причинно-следственной связи [Mittelstadt *et al.* 2016]. Этот подход порождает значительные проблемы в развитии прогностической аналитики.

Одна из наиболее ранних типологий видов предвзятостей, присущих алгоритмам (сами авторы рассматривали их как виды предвзятостей «компьютерных систем»), была опубликована еще в 1996 г. Б. Фридман и Х. Ниссенбаум [Friedman, Nissenbaum 1996]. Они выделили три типа предвзятостей:

- связанные с уже существовавшими на момент создания алгоритма реалиями;
- связанные с техническими ограничениями алгоритма;
- связанные с особенностями применения алгоритма на практике.

Предвзятости, связанные с существующими реалиями, могут не только быть имманентно присущи алгоритму, основанному, например, на глубоком обучении нейросети, но и целенаправленно внедряться разработчиками или операторами в дизайн системы, например при ручной корректировке индексов поисковых систем и критериев ранжирования. Более того, такая корректировка заставляет нас вновь обратиться к вопросу о влиянии ценностей и этических представлений разработчиков ИИ на возникновение предвзятостей в алгоритме, – хотя такое влияние может быть и совершенно непреднамеренным.

Техническая предвзятость, как в целом явствует из названия этого типа предвзятостей, возникает из-за технологических ограничений и/или ошибок – например, ошибка в конструкции генератора случайных чисел, из-за которой предпочтение отдается определенным числам. Аналогичным образом ошибки могут проявляться в наборах данных, обрабатываемых алгоритмами. Ошибки в данных непреднамеренно принимаются алгоритмом и скрываются в выходных данных и созданных моделях. Отчасти способствовать снижению частоты таких предвзятостей может мониторинг оператором-человеком – хотя в этом случае существует опасность возникновения других предвзятостей, связанных с человеческим вмешательством в алгоритм (описаны в предыдущем абзаце).

Возникающие (эмерджентные) предвзятости особенно характерны для интерфейсов; изменение контекста использования вполне может создать трудности для новой группы пользователей [Friedman, Nissenbaum 1996].

В принятой в ноябре 2021 г. «Рекомендации об этических аспектах искусственного интеллекта» ЮНЕСКО опасается «предвзятости, которую такие технологии (ИИ) способны порождать и усугублять, что потенциально может вести к дискриминации, неравенству, цифровому разрыву и маргинализации, ставить под угрозу культурное, социальное и биологическое разнообразие и усугублять социальное или экономическое расслоение»; подчеркивает, что необходимо обеспечение «прозрачности и понятности работы алгоритмов и данных, на основе которых проводится обучение интеллектуальных систем» и потенциальных последствий «их применения, в частности с точки зрения уважения человеческого достоинства, прав человека и основных свобод, гендерного равенства, демократии, участия в социально-экономических, политических и культурных процессах, научной и инженерной практики, защиты прав животных, а также состояния окружающей среды и экосистем» [ЮНЕСКО 2021].

Чрезвычайно тесно связана с проблемой алгоритмических предвзятостей проблема такого вида фильтрации информации, как персонализация. С одной стороны, она призвана снижать информационную нагрузку на пользователя (уменьшать объем информации, который должен воспринять пользователь для удовлетворения своих потребностей/желаний, за счет профилирования, в котором принимаются во внимание географическое положение и демографические характеристики пользователя, история предыдущих поисковых запросов и т. д.). С другой стороны, однако, персонализация может способствовать тому, что пользователь попадает в своего рода ловушку и не может получить ту информацию, которую фильтры, применяемые информационным сервисом, сочли неподходящей или нерелевантной для данного пользователя. При этом персонализация не только не устраняет имеющиеся в системе предвзятости, но и зачастую привносит новые [Bozdag 2013].

Существует точка зрения, что, поскольку алгоритм не может быть «виноват» в своей предвзятости, ответственность за неэтичные последствия его решений должны нести разработчики алгоритма:

«Я концептуализирую алгоритмы как ценностно нагруженные, а не нейтральные, поскольку алгоритмы создают моральные последствия, укрепляют или подрывают этические принципы, а также

обеспечивают или ослабляют права и достоинство заинтересованных сторон. Кроме того, алгоритмы играют важную роль в этических решениях и влияют на делегирование ролей и обязанностей в рамках этих решений. <…> Таким образом, фирмы, разрабатывающие алгоритмы, несут ответственность за определение того, насколько большую роль будет разрешено играть отдельным лицам в последующем алгоритмическом решении. Вопреки нынешним аргументам я считаю, что если алгоритм предназначен для того, чтобы лишить людей возможности брать на себя ответственность за принятие решения, то разработчик алгоритма должен нести ответственность за этические последствия используемого алгоритма» [Martin 2019: 835].

«Чтобы свести к минимуму предвзятость, дизайнеры должны представлять не только предполагаемую ситуацию использования системы, но и учитывать все более разнообразные социальные контексты использования. Затем проектировщики должны разумно предвидеть вероятные контексты их использования и дизайна. Если невозможно спроектировать для расширенных контекстов использования, проектировщики должны попытаться сформулировать ограничения на соответствующие контексты использования системы» [Bozdag 2013].

Однако при этом существует значительное число примеров того, как предвзятость алгоритмов ИИ не только и не столько становилась причиной несправедливых решений, дискриминирующих отдельные группы людей, обладающих тем или иным признаком, но скорее выполняла роль «лакмусовой бумажки», наглядно выявляющей дискриминирующий характер более ранних решений, принимавшихся людьми (поскольку эти решения попадали в набор данных, на котором обучался алгоритм). Здесь можно вспомнить истории крупнейших компаний мира – машинное обучение рекрутингового алгоритма компании Amazon привело к тому, что алгоритм стал отдавать предпочтение кандидатам-мужчинам перед кандидатами-женщинами, поскольку такой «перекос» имел место в реальной статистике найма компании, на которой обучался алгоритм. Алгоритм таргетирования рекламы Facebook позволял рекламодателям исключать определенные группы людей из числа тех, кому демонстрировалось то или иное рекламное объявление [больше примеров см.: Харитонова и др. 2021]. Иными словами, предвзятость реального мира в значительной степени формирует и предвзятость алгоритмов [Shaulova 2019].

Заключение

Предвзятость алгоритмов ИИ может быть частично ограничена стандартами и этическими принципами, закрепленными в программных документах (в качестве примера такого документа можно упомянуть, скажем, «Руководство по этике для надежного ИИ Специальной группы экспертов высокого уровня Совета Европы» – Ethics Guidelines for Trustworthy AI). Но одно лишь нормативное правовое регулирование неспособно устранить эту проблему. Во-первых, разнообразие принципов, рамок, руководств и стандартов создает возможность лавировать между ними, чтобы оправдать существующие алгоритмы «как есть», не внося в них изменения, а также создает проблему несопоставимости многих стандартов. Во-вторых, как было отмечено выше, предвзятость алгоритмов может в значительной степени порождаться предвзятыми решениями в реальной жизни, которые алгоритм «усваивает» во время обучения на реальных данных. Разработка механизмов выявления и корректировки этого «усвоения», таким образом, становится ключевым элементом в преодолении проблем неэтичности алгоритмов ИИ.

Литература

Железнов А. Мораль для искусственного интеллекта: перспективы философского переосмысления // Логос. 2021. Т. 31. № 6(145). С. 95–122.

Харитонова Ю. С., Савина В. С., Паньини Ф. Предвзятость алгоритмов искусственного интеллекта: вопросы этики и права // Вестник Пермского университета. Юридические науки. 2021. Вып. 53. С. 488–515.

Шиллер А. В. Место этической системы в архитектуре искусственного интеллекта // Вестник Томского государственного университета. 2020. № 456. С. 99–103.

ЮНЕСКО. Рекомендация об этических аспектах искусственного интеллекта. 2021 [Электронный ресурс]. URL: https://unesdoc.unesco.org/ark:/48223/pf0000380455_rus (дата обращения 11.02.2023).

Bostrom N. Superintelligence: Paths, Dangers, Strategies. Oxford : Oxford University Press, 2014.

Bostrom N., Yudkowsky E. The Ethics of Artificial Intelligence // Artificial Intelligence Safety and Security / ed. by R. V. Yampolskiy. New York : Routledge, 2018. Pp. 57–69.

Bozdag E. Bias in Algorithmic Filtering and Personalization // Ethics and Information Technology. 2013. Vol. 15. Pp. 209–227.

Friedman B., Nissenbaum H. Bias in Computer Systems // ACM Transactions on Information Systems (TOIS). 1996. Vol. 14. No. 3. Pp. 330–347.

Hawking, S., Russell, S., Tegmark, M., Wilczek, F. Transcendence Looks at the Implications of Artificial intelligence – But are We Taking AI Seriously Enough? [Электронный ресурс] : The Independent. 2014. May 1. URL: https://www.independent.co.uk/news/science/stephen-hawking-transcendence-looks-at-the-implications-of-artificial-intelligence-but-are-we-taking-ai-seriously-enough-9313474.html (дата обращения: 11.02.2023).

Macnish K. Unblinking Eyes: The Ethics of Automating Surveillance // Ethics and Information Technology. 2012. Vol. 14(2). Pp. 151–167.

Martin K. Ethical Implications and Accountability of Algorithms // Journal of Business Ethics. 2019. Vol. 160. Pp. 835–850.

Mittelstadt B. D., Allo P., Taddeo M., Wachter S., Floridi L. The Ethics of algorithms: Mapping the Debate // Big Data & Society. 2016. Vol. 3(2). 2053951716679679.

Müller V. C., Bostrom N. Future Progress in Artificial Intelligence: A Survey of Expert Opinion // Fundamental Issues of Artificial Intelligence / ed. by V. C. Müller. Berlin: Springer, 2016. Pp. 553–571.

Shaulova T. Artificial Intelligence vs. Gender Equality // International Relations and Dialogue of Cultures. 2019. Vol. 7. Pp. 52–54.

Siau K., Wang W. Building Trust in Artificial Intelligence, Machine Learning, and Robotics // Cutter Business Technology Journal. 2018. Vol. 31. No. 2. Pp. 47–53.