УДК 17.026 Вестник СПбГУ. Философия и конфликтология. 2024. Т. 40. Вып. 3

Может ли искусственный интеллект быть ксенофобом? Конфликтные и этические вызовы развития нейросетей*

А. Н. Сунами

Санкт-Петербургский государственный университет,
Российская Федерация, 199034, Санкт-Петербург, Университетская наб., 7–9

Для цитирования: *Сунами А. Н.* Может ли искусственный интеллект быть ксенофобом? Конфликтные и этические вызовы развития нейросетей // Вестник Санкт-Петербургского университета. Философия и конфликтология. 2024. Т. 40. Вып. 3. С. 459–472.
https://doi.org/10.21638/spbu17.2024.308

В статье исследуются конфликтные и этические вызовы, связанные с «проблемой предвзятости» нейросетей. Опираясь на попытки концептуализации проблемы предвзятости цифровых алгоритмов Кейт Кроуфорд, Иезекииля Диксон-Романа и Лучаны Паризи, комплекс исследований естественного языка для программирования, теорию культурного насилия Йохана Галтунга, работы Джона Ролза и Уилла Кимлики о проблеме справедливости и неравенства, автор приходит к выводу, что опыт рефлексии цифровых алгоритмов, построенных на технологии нейронных сетей, позволил к настоящему времени сформировать корпус этически окрашенных проблем, центральной из которых является «проблема предвзятости». Эта проблема в свою очередь непосредственно связана с феноменом «естественного языка», т. е. отражением языковых практик прошлого и настоящего, срезом оцифрованного опыта человечества, который является основным ресурсом обучения нейросетевых алгоритмов. В этом контексте особое внимание уделяется стереотипам, властным иерархиям, неравенствам и дискриминации, закодированным в естественном языке, которые могут быть описаны в терминах политики классификации и культурного насилия. Более того, монополизация информационного пространства глобальными игроками усиливает эти эффекты, приводя к отражению в работе нейросетей проектируемого в качестве доминирующего определенного социального дискурса. В статье показывается, что этические и конфликтные последствия «проблемы предвзятости» могут быть сокращены только при условии получения той или иной социальной группой статуса дискриминируемой. На основе выделенных базовых и субстанциональных характеристик текущих нейросетевых алгоритмов автор приходит к выводу о высокой опасности индоктринации алгоритмов искусственного интеллекта элементами ксенофобии, исключительности и «культуры отмены».

Ключевые слова: цифровизация, нейросеть, проблема предвзятости, естественный язык, неравенство, политика классификации, культурное насилие.

Выступая в ноябре 2023 г. на пленарном заседании Международной конференции по искусственному интеллекту и машинному обучению «Artificial Intelligence Journey 2023» на тему «Революция генеративного ИИ: новые возможности», Прези-

* Исследование выполнено за счет гранта Российского научного фонда № 23-28-00220, https://rscf.ru/project/23-28-00220/.

дент России Владимир Путин заявил следующее: «…перед машиной ставят какую-то творческую задачу, и она решает ее, используя только англоязычный массив данных, тот, который удобен, выгоден разработчикам системы. Таким образом, алгоритм, например, может указать машине, что России, нашей культуры, науки, музыки, литературы просто не существует. Своего рода “отмена” в цифровом пространстве. А потом так же могут поступить и с другими культурами и другими цивилизациями, выпячивая себя, подчеркивая и в этом пространстве свою исключительность»[1]. «Вот такой ксенофоб может получиться из искусственного интеллекта, созданного по некоторым западным стандартам и лекалам», — резюмировал Президент.

Как нами уже приходилось писать ранее, цифровизация жизни ежедневно ставит перед академическим сообществом новые вызовы [1]. Переходная фаза от экзотики инновации до повседневного использования драматически сокращается, рискуя однажды окончательно схлопнуться. Вынужденно признавая, что моральные философы, которые стараются запечатлеть, как правило, менее подвижные объекты, не всегда успевают совершить глубокую и проработанную этическую рефлексию новых цифровых феноменов, мы, тем не менее, полагаем небесполезным попытаться набросать эскиз моральных вызовов, которые уже не стучатся в дверь, а переступили порог.

Цитата из выступления Владимира Путина, с которой мы начали настоящую статью, фиксирует крайне неприятный факт того, что внедрение в повседневную практику технологий, построенных на нейросетях и машинном обучении, делает прежде аморфное верифицируемым. И действительно, десятилетие назад национальное присутствие в общемировом культурном пространстве, его значительность, объем и прочие параметры, отражающие успехи в использовании мягкой силы, носили скорее умозрительный и манипулятивный характер и бытовали зачастую в форме риторических вопросов: что такое современный мировой театр без Антона Чехова, а кино без Константина Станиславского?.. Теперь же, в силу того что все дискурсивное пространство становится ресурсом для алгоритмов искусственного интеллекта, оказалось, что его можно строго математически калькулировать. Иными словами, некий валовый культурный импакт становится вычисляемым и могущим быть подвергнутым сравнению с прочими конкурирующими образами. Одновременно с этим перспективы изоляции какой-либо культуры, замыкания ее носителя на самом себе в некой резервации, становятся конкретными и выполняемыми на практике.

Но прежде чем мы приступим к оценке этой перспективы и анализу того, насколько многополярность и плюрализм являются включенными в качестве базовых характеристик в формируемую цифровую реальность, необходимо совершить небольшой экскурс в историю вопроса, откуда взялись нейросети и каковы ключевые параметры их работы.

Современные энциклопедии достигли определенного консенсуса в том, кто явился пионерами в постановке и решении этой интеллектуальной задачи. Считается, что исторически первой была математическая модель нейрона Уоррена Мак-Каллока (Warren McCulloch) и Уолтера Питтса (Walter Pitts), построенная в 1943 г., в которой авторы формализовали понятие нейронной сети. В своей работе исследо-

[1] Конференция «Путешествие в мир искусственного интеллекта» // Президент России. 2023. URL: https:// http://www.kremlin.ru/events/president/news/72811 (дата обращения: 01.12.2023).

ватели опирались на следующую гипотезу: «Из-за принципа нервной деятельности “все или ничего” нейронные события и отношения между ними могут трактоваться с помощью препозиционной логики» [2, с. 115]. Таким образом, модель поведения любой «живой» нейронной сети может быть описана в логико-математических терминах и, соответственно, искусственно воспроизведена программными средствами. Итак, базовый подход к определению нейросети с этого момента сформировался в виде следующего тезиса: нейронная сеть — это математическая модель сама по себе или же ее программная реализация, построенная на принципах аналогии с нейронными клетками биологического организма и их сетями. Необходимо отметить, что 1940-е годы оказались чрезвычайно плодотворными в плане развития теории нейронных сетей. Помимо построения модели Мак-Каллока — Питтса необходимо отметить еще два важных события. Во-первых, в 1948 г. выходит фундаментальная работа американского математика и одного из основоположников кибернетики Норберта Винера (Norbert Wiener) «Кибернетика, или Управление и связь в животном и машине», не только ставшая краеугольным камнем всех последующих работ об искусственном интеллекте, но и получившая невероятный отклик у простых читателей; в ней впервые в академическом ключе был использован термин «кибернетика», проведено масштабное сравнительное исследование работы мозга и вычислительной машины и описаны вероятные возможности самообучение и саморазвития компьютерной программы на рефлексии ею своих ошибок [3]. Во-вторых, в 1949 г. канадский физиолог Дональд Хебб (Donald Hebb) формулирует свое знаменитое правило, гласящее: «Если аксон клетки А находится достаточно близко, чтобы возбуждать клетку В, и многократно либо постоянно участвует в ее возбуждении, то наблюдается определенный процесс роста или метаболических изменений в одной либо обеих клетках, способствующий усилению эффективности клетки А как одного из активаторов клетки В» [4, p. 62]. Иными словами, связь между нейронами совершенствуется по мере увеличение частоты использования. В начале следующего десятилетия была осуществлена первая успешная попытка реализовать на практике предшествующие теоретические выкладки, когда в 1951 г. под руководством Марвина Минского (Marvin Minsky), будущего сооснователя Лаборатории информатики искусственного интеллекта Массачусетского технологического института, была сконструирована первая обучающаяся машина со случайно связанной нейросетью SNARC, состоящей из 40 синапсов, работающая на принципе Хебба.

С этого момента в течение последующих десятилетий предпринимались попытки, с одной стороны, создания все более совершенных нейронных сетей, как то перцептрон Розенблатта и сконструированный на его основе первый нейрокомпьютер Mark I, с другой — разрешения задач вроде парадокса Моравека, утверждающего, что «относительно легко достичь уровня взрослого человека в таких задачах, как тест на интеллект, или игре в шашки, однако сложно или невозможно достичь навыков годовалого ребенка в задачах восприятия или мобильности» [5, p. 15]. Без прорыва в этих направлениях дальнейшее поступательное развитие не было бы возможным. Решение этих проблем в совокупности позволило к настоящему времени выйти на принципиально иной уровень функционирования нейросетей, сделало возможным включение их не только в решение серьезных академических задач, например, предсказания пространственной структуры белка с помощью нейросети AlpfaFold [6] или управленческих проблем [7], но и в по-

вседневную практику обычных людей. Нейросети, способные генерировать тексты или компьютерные программы, отвечать на вопросы пользователей, как, например, ChatGPT (Generative Pre-trained Transformer) — чат-бот с генеративным искусственным интеллектом от компании OpenAI, или YandexGPT от российской компании Яндекс, внедренный в помощника «Алису», активно используются обычными людьми в решении бытовых вопросов, рабочих задач или в развлекательных целях. Не стоит удивляться тому, что в России на правительственном уровне всерьез обсуждается проблема использования студентами нейросетей в написании квалификационных работ, а Северный (Арктический) федеральный университет первым в России нормативно зафиксировал, какая доля текста, написанного таким образом, считается допустимой в работе, фактически легализовав для своих обучающихся использование искусственного интеллекта.

Каковы же стандартная архитектура нейросети и принципы ее работы? Попробуем сгенерировать ответ на этот вопрос с помощью YandexGPT 2.0. Ответ готов через доли секунды: «Нейросети, также известные как искусственные нейронные сети, основаны на принципах работы биологических нейронов. Они представляют собой систему соединенных между собой слоев искусственных нейронов, которые обрабатывают информацию и принимают решения. <…> Нейросети обучаются на наборах данных, используя алгоритмы обучения, такие как обратное распространение ошибки или метод стохастического градиентного спуска. Эти алгоритмы корректируют веса связей между нейронами на основе разницы между ожидаемым и фактическим выходом сети для каждого обучающего примера. Количество скрытых слоев и нейронов в них, а также тип функции активации определяют сложность и возможности нейросети. Более сложные сети с большим количеством слоев и нейронов могут обрабатывать более сложные и абстрактные представления данных, но также требуют больших вычислительных ресурсов и времени на обучение. В общем, нейросети работают путем последовательного применения функций активации на каждом слое сети, что позволяет им аппроксимировать сложные зависимости и взаимосвязи в данных».

Таким образом, не прилагая большого труда, мы получаем сформулированный искусственным интеллектом вполне себе корректный набор базовых принципов работы нейронных сетей. Итак, суммируя описанное выше, мы можем зафиксировать, что современные нейронные сети — это алгоритм машинного обучения, работающий по схеме, схожей с работой головного мозга. Его характерными чертами можно назвать:

— архитектуру (нейросеть состоит из одного или нескольких слоев нейронов, объединенных в определенной последовательности);

— взаимодействие между нейронами (нейроны в каждом слое обмениваются сигналами друг с другом, каждый из них принимает входные сигналы, выполняет некоторые вычисления и передает выходной сигнал другим нейронам);

— веса и смещения (веса и смещения определяют важность каждого входного сигнала в нейронах, способны подстраиваться, чтобы минимизировать возможные ошибки);

— функции активации (функции активации нейронов определяют, какой выходной сигнал будет передан следующему слою, такая нелинейность позволяет решать сложные задачи);

— обучение (нейросеть обучается на наборе данных, где для каждого входа есть правильный выход, в результате происходит оптимизация веса и смещения);

— инференцию (после обучения нейросеть может использоваться для предсказания новых данных).

Мы не будем подробно останавливаться на методологии обучения нейросетей, ибо к текущему моменту она достаточно полно разобрана в научной литературе [8]. Для настоящей темы самым важным является вопрос, что выступает источником первичных данных для такого обучения. В зависимости от направленности нейросети это могут быть специально подготовленные под конкретную задачу обучающие наборы данных, открытые датасеты, содержащие пакеты изображений или текстов, а также массив данных из интернета. Так, например, в докладе, посвященном описанию работы языковой модели GPT-3, представленной OpenAI в мае 2020 г., показано, что ресурсами для обучения выступили несколько датасетов (в процентах указаны пропорции каждого источника в миксе данных для обучения): открытый репозиторий веб-данных Common Crawl (60 %), тексты вебстраниц, заслуживших одобрение пользователями Reddit (22 %), два корпуса книг, размещенных в интернете (16 %), и статьи англоязычной википедии (3 %) [9, p. 8]. Таким образом, можно заметить, что использованный для обучения набор данных представляет собой контент с невысоким порогом доступа, близкий по своему характеру к реальной коммуникации людей в пространстве интернета. Здесь стоит отметить деталь, которая чрезвычайно важна для анализа потенциальной ксенофобии нейросетей, — проблему естественного языка, или программирования на естественном языке (natural-language programming (NLP)).

Проблематика естественного языка, т. е. языка, используемого людьми для общения, является одной из важнейших в лингвистике и философии языка. Не менее значима она в настоящее время для программирования. Анализ и работа с естественными языками — ключевая задача для разработчиков, которые стремятся создать машины, способные обрабатывать и генерировать информацию на уровне, близком к человеку. Подразделом программирования на естественном языке является понимание естественного языка (natural language understanding, NLU) — комплекс задач, среди прочего включающий в себя логический вывод на естественном языке (natural language inference), ответы на вопросы и понимание прочитанного. Основная характеристика задач понимания машиной естественного языка заключается в том, что эти задачи чрезвычайно комплексны и обычно «требуют знаний о мире и здравого смысла» [10, p. 110]. На последней позиции необходимо остановится подробнее. Людмила Яковлева в объемной статье, посвященной традиции философской рефлексии здравого смысла, резюмирует, что «философская традиция, начатая Сократом, продолженная Платоном и Аристотелем, поддержанная стоиками и Цицероном, далее переданная по эстафете через средневековых философов — Тертуллиана, Августина, мистиков к Декарту и другим представителям XVII–XVIII вв., определила, что для совместного познания и жития люди должны обладать общими для них первоначальными знаниями, ценностями, нормами, в коих они не должны сомневаться» [11, с. 47], а чуть выше замечает многозначность этого термина и трудности его перевода: «“common sense” <…> можно перевести одним словосочетанием: “здравый смысл”, а можно дать прямой перевод: “общее чувство”, “хорошее чувство”, “правильное мышление”» [11, с. 31]. Все это рождает

довольно-таки неоднозначное восприятие это феномена. Философская традиция скорее нацеливает нас на отделение здравого смысла от обыденного сознания. В то время как менее строгий подход, вероятнее всего, представляет здравый смысл как некое среднее мнение людей определенной локации и времени, детерминируемое, кроме собственно разума, предрассудками, стереотипами, пристрастиями, привычками, интересами, опытом и прочими подобными переменными. Так, зачастую в социологической литературе здравый смысл полагается предшествующим общественному мнению. Мы можем предположить, что в случае с разбираемой нами проблемой менее строгий подход является гораздо более уместным, фактически уравнивая здравый смысл и общепринятое понимание, которое, в свою очередь, является отражением того дискурсивного поля, которое предоставляет ресурсы для обучения машины. Эти дискурсивные поля, пользуясь терминологией Мишеля Фуко, отражают властные отношения и укоренившиеся идеологические форматы [12]. К этой мысли мы вернемся позднее. Таким образом, вполне закономерно, что чем ближе разработчик нейросетевой программы подходит к решению проблемы естественного языка, тем конкретнее в результатах ее работы будут отражаться черты, характерные для текущего социального дискурса в целом или какого-либо его сегмента, в том числе и те неприглядные элементы, от которых мы стараемся дистанцироваться, делая вид, что их не существует.

Одной из наиболее существенных проблем для разработчиков рассматриваемого программного продукта является «проблема предвзятости» (the trouble with bias). Проблема столь серьезна, что теперь фактически каждая команда программистов с необходимостью должна иметь в штате так называемого специалиста по «анализу этических последствий». В докладе по GPT-3, к которому мы уже обращались выше, этой проблеме уделяется значительное место, и формулируется она следующим образом: «…предубеждения, присутствующие в обучающих данных, могут привести к тому, что модели будут генерировать стереотипный или предвзятый контент <…> это вызывает беспокойство, поскольку предвзятость моделей может по-разному навредить людям в соответствующих группах, закрепляя существующие стереотипы» [9, с. 38].

Тема неравенства в самых разнообразных пространствах и конструкциях, невероятно востребованная в современной академической литературе, к настоящему времени начинает все больше волновать исследователей в контексте того, как неравенство отражается в естественном языке и как оно влияет на программирование на естественном языке. Например, именно под таким углом анализируется проблема гендерного неравенства в английской литературе XIX — начала XX в. в свежей работе ученых Университета Южной Калифорнии, вышедшей в январе 2024 г. В результате исследования, проведенного на выборке из 2443 художественных текстов, срок действия авторских прав на которые истек, посредством воспроизводимой методологии data science, основанной на базовых компонентах программирования на естественном языке, были извлечены различные показатели распространенности гендерных признаков. Это позволило «убедительно продемонстрировать значительное несоответствие в пропорции женских и мужских персонажей <…> а описания, связанные с женскими персонажами по всему корпусу, заметно стереотипны в сравнении с описаниями, связанными с мужскими персонажами» [13, p. 77]. Иными словами, гендерное неравенство выражается не только в количе-

ственных, но и в качественных параметрах, демонстрируя шаблонность женских образов, выполняющих скорее функцию фона для основных действующих лиц. Здесь, мы можем повторить тезис, озвученный в начале. Если ранее можно было в дискуссии о неравенстве использовать в качестве аргумента отсылки к детально прописанным женским образам, настолько вошедшим в сами несущие конструкции мировой культуры, что без них она является непредставимой, то теперь в силу возможностей квантификации обширного корпуса данных, подобного рода «качественные» аргументы дезавуируются. Здесь необходимо обратить внимание на еще одну важную деталь программирования на естественном языке. Дело в том, что используемый для обучения машины естественный язык нельзя полагать полностью синхронизированным с текущими языковыми практиками. Корпус данных, которыми оперируют разработчики программного продукта, в качестве которого может быть использована вся совокупность оцифрованных текстов, изображений и других категорий кодированной информации, выходит за пределы новейшего периода, когда «инклюзивный» язык, т. е. избегание слов и действий, оскорбляющих отдельные группы населения, иными словами, языка вражды, начал активно внедряться, иногда принудительными мерами, в те или иные дискурсивные практики. Мы должны отдавать себе отчет в том, что табуированные в настоящее время языковые конструкции не так давно широко использовались, более того, были неотъемлемой частью «нормального» языка, как в повседневном общении, так и в формализованной коммуникации (например, в политических текстах). Иногда в текущих обстоятельствах это действительно может приводить к манифестации требований «отмены» того или иного классика или общественного деятеля, не вписывающегося в прокрустово ложе современного политкорректного языка. Аналогичное затруднение возникает при попытке экстраполировать текущие и, что немаловажно, юридически значимые стандарты разжигания розни или пропаганды наркотиков на произведения литературы или изобразительного искусства, созданные в эпоху, когда этих стандартов еще не существовало в принципе или же они носили крайне размытый характер. Таким образом, топливом для естественного языка нейросети является весь оцифрованный человеческий опыт, со всеми его изъянами. Это касается не только всечеловеческого культурного багажа, но и опыта отдельных корпораций. Иллюстрацией этому является ставший уже трюизмом скандал, связанный с внедрением гигантом мировой торговли Amazon алгоритма подбора соискателей вакантных должностей, основанного на технологиях искусственного интеллекта. Нейросеть обучалась на выборке резюме всех сотрудников за последние десять лет, в которой мужчин оказалось больше, чем женщин, в силу чего искусственный интеллект пришел к выводу о предпочтительности кандидатов-мужчин. Полагая, что даже коррекция гендерных параметров не решит проблему, ибо, основываясь на предыдущем опыте, алгоритм может дискриминировать кандидатов по другим признакам, компания отказалась от использования программных средств при найме на работу. Можно смело предположить, что урон репутации был нанесен не только в связи с некорректной работой нейросети, которую можно было бы легко списать на несовершенство разработки, но и тем, что алгоритм фактически сбросил маску с ранее латентных дискриминационных практик.

Возвращаясь к «проблеме предвзятости», необходимо отметить, что в последних работах на эту тему акценты смещаются в сторону ее рассмотрения через опти-

ку более широкого контекста. Так, например, известный специалист в области искусственного интеллекта и общественный деятель Кейт Кроуфорд (Kate Crawford), старший научный сотрудник Microsoft и профессор Калифорнийского университета, в своем бестселлере 2021 г. «Atlas of AI. Power, Politics, and the Planetary Costs of Artificial Intelligence», переведенном в 2023 г. на русский язык («Атлас искусственного интеллекта: руководство для будущего»), предпочитает использовать термин «политика классификации».

Проблему классификаций она пытается раскрыть через отсылку к деятельности известного американского натуралиста и краниолога первой половины XIX в. Сэмюэля Мортона (Samuel Morton), собравшего огромную коллекцию черепов и занимавшегося их сортировкой по расовому признаку. Многие выводы, сделанные им исходя из этой квалификации, были результатом манипуляций и подмен с целью подтверждения тезиса о превосходстве одних рас над другими. Важным является то, что, по мнению Стивена Гулда (Stephen Jay Gould), чью цитату приводит в своей книге Кейт Кроуфорд, нет «никаких доказательств сознательного мошенничества» [14, p. 88] со стороны Сэмюэля Мортона, скорее мы имеем дело с частным случаем известного когнитивного искажения — склонностью к подтверждению своей точки зрения (confirmation bias), т. е. стремлением человека искать или тенденциозным образом интерпретировать информацию в соответствии с предустановленными убеждениями или гипотезами. Иными словами, Сэмюэль Мортон был настолько и искренне убежден в превосходстве белой расы, что черепа бессознательно отбирались, сортировались и сами собой укладывались в предначертанное построение. Кейт Кроуфорд ставит проблему следующим образом: «Хотя и существует тенденция сосредоточиться на ошибках в измерениях черепа и на том, как их исправить, гораздо бóльшая проблема кроется в мировоззрении, лежащем в основе этой методологии. Таким образом, цель должна состоять не в том, чтобы призвать к более точным или “справедливым” измерениям черепа для укрепления расистских моделей интеллекта, а в том, чтобы полностью осудить этот подход» [15, с. 120]. Классификация, с точки зрения Кейт Кроуфорд, — это «акт власти», потому и уместно ее описывать в политических терминах. Кроме очевидных параллелей с Мишелем Фуко, необходимо вспомнить также Йохана Галтунга (Johan Galtung), чей известный и оригинальный концепт «культурного насилия» вполне может быть рассмотрен в этой связи. Под культурным насилием он подразумевает «те аспекты культуры, символической сферы нашего существования, примером которых являются религия и идеология, язык и искусство, эмпирическая наука и формальная наука (логика, математика), которые могут быть использованы для оправдания или узаконивания прямого или структурного насилия» [16, p. 291]. Вероятно, для Кейт Кроуфорд классификации являются феноменом, вполне рифмующимся с культурным насилием, так как работают по такому же принципу и со схожей целью: «…схемы социального упорядочивания натурализуют иерархии и увеличивают неравенство» [15, с. 121]. Более того, приводимая ею цитата Джеффри Боукера (Geoffrey C. Bowker) и Сьюзен Ли Стар (Susan Leigh Star), согласно которой классификации могут исчезать «в инфраструктуре, в привычке, в чем-то само собой разумеющемся» [17, p. 319], также позволяет провести параллель со структурным и культурным насилием, которые бóльшую часть времени бытуют как ординарный фон социальной жизни и не осознаются их жертвами. Как справедливо замечает Кейт Кроуфорд, еще десятилетие назад мнение о предвзятости искусствен-

ного интеллекта выглядело как попытка соригинальничать, в настоящее же время многочисленными примерами дискриминационных систем вряд ли кого-то можно удивить. Вывод очевиден: если искусственный интеллект использует классификации для кодирования и закрепления власти, или, иными словами, иерархии и неравенства, радикальным решением проблемы будет исключение классификаций из уравнения [15, с. 121–122].

В схожем ключе описывают проблему предвзятости искусственного интеллекта американские исследователи Иезекииль Диксон-Роман (Ezekiel Dixon-Roman) и Лучана Паризи (Luciana Parisi). Однако их подход предполагает несколько иной фокус, чем у Кейт Кроуфорд. Теоретические построения последней, несмотря на то что в классификациях могут быть обнаружены универсальные черты, все-таки оставляют так или иначе место для определенного культурного плюрализма. Кажется, имеет право быть высказанным допущение, что классификации могут быть описаны в терминах того или иного автономного пространства, ибо очевидно, что расовая классификация для России имеет несравненно меньший вес в кодировании власти, чем, например, в Соединенных Штатах. В работе же Иезекииля Диксон-Романа и Лучаны Паризи проблема предвзятости поднимается на наднациональный уровень, становясь отражением глобального постколониального капитализма. Развивая положения «социогенного принципа» Франца Фанона (Frantz Fanon) и Сильвии Винтер (Sylvia Wynter), т. е. объяснение того, как социально-политическое обретает телесность, они утверждают, что «автопоэтический институт социогенного кода пронизывает не только человеческие онтологии, но и более чем человеческие онтологии, включая социотехнические совокупности данных и алгоритмов» [18, p. 118]. В своих работах Сильвия Винтер идентифицирует эту социально-политическую сборку человека как «западного человека» и утверждает, что актуальная и доминирующая версия «происхождения мира» соответствует биогуманистическому типу homo oeconomicus, для которой корреляция между биологическим и экономическим выживанием определяет эпистемологическое объяснение того, кто успешен как биологический вид, а кто нет [19]. Отталкиваясь от этой позиции, Иезекииль Диксон-Роман и Лучана Паризи утверждают, что «социогенное кодирование другого как негативного маркера необходимо для рекурсивных циклов колониального предпринимательства, посредством чего естественная диадическая структура эквивалентности между человеком и миром гарантирует, что все остается неизменным под солнцем Запада <…> требование расширить нормализованную категорию человека, включив в нее исключенные различия и пролив свет на слепоту машины, не обеспечивает беспрепятственного политического переворота» [18, p. 118]. Здесь необходимо поставить еще один вопрос: в какой мере естественный язык свободен от искусственного вмешательства, но к размышлению о нем мы вернемся чуть позднее. Таким образом, вывод, к которому приходят авторы, близок к заключению Кейт Кроуфорд: попытки косметически отредактировать отдельные положения обученной на естественном языке машины посредством специалистов по «анализу этических последствий» будут напоминать попытки ковшиком вычерпать воду из получившего пробоину многотонного лайнера. Но даже эта борьба с ветряными мельницами требует еще одной процедуры, которую крайне важно упомянуть.

Здесь мы должны обратить внимание на связь предвзятости и социального конфликта. И в данном случае мы имеем в виду неочевидный факт, что предвзя-

тость отражает не равенство, а скорее конфликтный характер процесса признания предвзятости и получение той или иной группой статуса дискриминируемой как необходимого условия включения конкретной предвзятости в пул проблем, требующих внимания со стороны отделов анализа этических последствий. Совершенно очевидно, что статус дискриминируемой группы является приобретаемым. Точнее, можно сформулировать этот тезис следующим образом: лишь некоторые группы в настоящее время имеют такой статус, при этом реальное число дискриминируемых сообществ, скорее всего, гораздо больше. Какова же механика включения или невключения в данный круг? Первый аспект — основания. Современной дискриминации меньшинств и этико-политическим аспектам сглаживания неравенств посвящены ставшие уже классическими работы политических философов Роберта Нозика (Robert Nozick), Джона Ролза (John Rawls), Уилла Кимлики (Will Kymlicka), Майкла Уолцера (Michael Walzer) и др. Мы не будем вдаваться здесь в подробности их размышлений, но центральным можно считать вопрос: какое неравенство будет справедливым или естественным, а какое нет? Уилл Кимлика обращается к известному положению теории справедливости Джона Ролза: «Все социальные ценности — свобода и благоприятные возможности, доходы и богатство, социальные основы самоуважения — все это должно быть равно распределено, кроме тех случаев, когда неравное распределение любой, или всех, из этих ценностей дает преимущество каждому» [20, с. 67]. При этом Уилл Кимлика интерпретирует этот тезис следующим образом: «Мы относимся к людям как к равным, если устраняем не все неравенства, а только те, которые причиняют кому-то ущерб. <…> Неравенства допустимы, если они улучшают мою, равную с другими долю, но они недопустимы, если посягают на долю, которая полагается мне по справедливости» [21, с. 81]. Еще одним основанием, согласно подходу Уилла Кимлики, является происхождение неравенства. Если причина неравенства не зависит от воли неравного, как то гендер или раса, справедливым будет его устранить. Если же неравенство является результатом выбора или иных действий, оно является справедливым. Второй аспект — механизм признания. Полагаем, не будет смелым предположение, что конфликт является основным способом получения статуса дискриминируемой группы. Ведь именно борьба за права, за расширение общественного договора, в конечном счете привела к признанию целого ряда общностей как нуждающихся в особом внимании. Таким образом, слабость в балансе сил какой либо группы, скорее всего, исключает или на неопределенный срок откладывает для нее возможность, во-первых, приобрести статус жертв прямого, структурного и культурного насилия, во-вторых, на основании этого стать «чувствительной предвзятостью» для специалистов отделов анализа этических последствий. Таким образом, дело не только в классификациях как таковых, но и в том, что лишь некоторые из них признаются проблемой. В этой связи необходимо заметить, что наблюдаемая сейчас в европейском пространстве кампания стигматизации по признаку российского гражданства, подобную которой сложно вспомнить за последнее время, ибо даже в разгар борьбы с международным терроризмом подозрительность в отношении всех мусульман редко когда носила столь манифестный характер, ярко иллюстрирует вышеприведенное положение и демонстрирует невысокие шансы россиян попасть в круг «проблем предвзятости».

Теперь мы можем вернуться к анонсированному выше вопросу, который будет заключительной из рассмотренных нами проблем. Само словосочетание «естественный язык» может стать причиной определенного заблуждения, которое заключается в иллюзии, что он является исключительно стихийно формируемым срезом некоей языковой практики. Образно выражаясь, результатом жестов «невидимой руки» интернета. Между тем, как мы уже писали в наших работах, не столь давно отгремевшие политические скандалы, причиной которых были доказанные случаи намеренной политической цензуры со стороны глобальных цифровых компаний, искажение ими алгоритмов в пользу определенных властных групп, показали, что «невидимой руки» интернета на самом деле не существует [22]. Безусловно, будет крайне сложно детально посчитать масштаб искажения, но сам факт его наличия вряд ли можно будет оспорить. Место свободных самоуправляемых ассоциаций заняли IT-гиганты, точнее, они давно уже там были, только странным образом не воспринимались пользователями в качестве игроков, имеющих свои интересы, и не всегда открытых бенефициаров. Таким образом, опасность монополизации естественного языка, которую мы пока не готовы обсуждать подробно, является еще одной угрозой многополярности и плюрализму формируемой цифровой реальности.

Итак, проведенный анализ позволяет нам сделать несколько выводов. Во-первых, экспоненциальное развитие нейросетей за последнее десятилетие позволило обобщить имеющийся опыт этических издержек их реальной работы и сформулировать целый ряд проблем, важнейшей из которых является «проблема предвзятости». Во вторых, «проблема предвзятости», очевидно, ассоциирована с феноменом «естественного языка», который является главным ресурсом машинного обучения. В свою очередь, естественный язык есть отражение текущих и прошлых языковых практик, всего оцифрованного опыта, в том числе включающего стереотипы и другие феномены «политики классификации» или «культурного насилия». В третьих, «проблема предвзятости» может быть применена только к тем группам, которые имеют «чувствительный» статус дискриминируемых, приобретаемый в результате конфликта. Слабая позиция какой-либо группы в балансе сил не позволяет ей получить маркировку дискриминируемой и попасть в пул «проблемы предвзятости». Наконец, в четвертых, монополизация информационного пространства глобальными игроками приводит к искусственному искажению естественного языка и отражению в работе нейросетей социального дискурса, проектируемого в качестве доминирующего. Все это позволяет сделать заключение о высоком уровне опасности индоктринации алгоритмов искусственного интеллекта элементами ксенофобии, исключительности и «политики отмены».

Литература

1. Сунами, А. Н. (2023), Этика «цифрового общества»: новый конфликт или новый баланс, *Вестник Санкт-Петербургского университета. Философия и конфликтология*, т. 39, вып. 3, с. 544–556. https://doi.org/10.21638/spbu17.2023.311

2. McCulloch, W. S. and Pitts, W. (1943), A logical calculus of the ideas immanent in nervous activity, *The Bulletin of Mathematical Biophysics. Kluwer Academic Publishers,* vol. 5, no. 4, pp. 115–133. https://doi.org/10.1007/BF02478259

3. Wiener, N. (1948), *Cybernetics or Control and Communication in the Animal and the Machine*, New York: Technology Press.

4. Hebb, D. O. (1949), *The Organization of Behavior. A Neuropsychological Theory*, New York: John Wiley & Sons.

5. Moravec, H. (1990), *Mind Children. The Future of Robot and Human Intelligence*, Harvard: Harvard University Press.

6. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman,D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A. W., Kavukcuoglu, K., Kohli, P. and Hassabis, D. (2021), Highly accurate protein structure prediction with AlphaFold, *Nature*, vol. 596, pp. 583–589. https://doi.org/10.1038/s41586-021-03819-2

7. Лисовский, А. Л. (2020), Применение нейросетевых технологий для разработки систем управления, *Стратегические решения и риск-менеджмент*, т. 11, № 4, с. 378–389. https://doi.org/10.17747/2618-947X-2020-4-378-389

8. Созыкин, А. В. (2017), Обзор методов обучения глубоких нейронных сетей, *Вестник ЮУрГУ. Серия: Вычислительная математика и информатика*, т. 6, № 3, с. 28–59. https://doi.org/10.14529/cmse170303

9. Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, A. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I. and Amodei, D. (2020), *Language Models are Few-Shot Learners*. https://doi.org/10.48550/arXiv.2005.14165

10. Du, M., He, F., Zou, N., Tao, D. and Hu, X. (2023), Shortcut Learning of Large Language Models in Natural Language Understanding, *Communications of the ACM,* vol. 67, no. 1, pp. 110–120. https://doi.org/10.1145/3596490

11. Яковлева, Л. И. (2007), Понятие здравого смысла и традиция, его конституирующая, *Вестник Московского университета. Серия 7: Философия*, вып. 4, с. 29–49.

12. Фуко, М. (1996), *Воля к истине: по ту сторону знания, власти и сексуальности. Работы разных лет,* М.: Касталь.

13. Kejriwal, M. and Nagaraj, A. (2024), Quantifying Gender Disparity in Pre-Modern English Literature using Natural Language Processing, *Journal of Data Science*, vol. 22, no. 1, pp. 77–96. https://doi.org/10.6339/23-JDS1100

14. Gould, S. J. (1996), *The Mismeasure of Man*, New York: W. W. Norton & Co.

15. Кроуфорд, К. (2023), *Атлас искусственного интеллекта: руководство для будущего*, М.: АСТ.

16. Galtung, J. (1990), Cultural Violence, *Journal of Peace Research*, vol. 27, no. 3, pp. 291–305.

17. Bowker, G. C. and Star, S. L. (1999), *Sorting Things Out: Classification and Its Consequences,* Cambridge, Mass.: MIT Press.

18. Dixon-Roman, E. and Parisi, L. (2020), Data capitalism and the counter futures of ethics in artificial intelligence. *Communication and the Public*, vol. 5, no. 3–4, pp. 116–121. https://doi.org/10.1177/2057047320972029

19. Wynter, S. (2001), Towards the sociogenic principle: Fanon, identity, the puzzle of conscious experience, and what it is like to be Black, in: Gomez-Moriana, A. and Duran-Cogan, M. (eds), *National identities and sociopolitical changes in Latin America*, Milton Pak: Routledge, pp. 30–66.

20. Ролз, Д. (1995), *Теория справедливости*, Новосибирск: Изд-во Новосиб. гос. ун-та.

21. Кимлика, У. (2010), *Современная политическая философия: введение*, М.: Высшая школа экономики.

22. Сунами, А. Н. и Павлова, Е. В. (2024), Моральные основания политики в условиях эластичного цифрового общества: перестанут ли ценности иметь значение? (часть 1), *Политическая экспертиза: ПОЛИТЭКС*, т. 20, № 1, с. 33–44. https://doi.org/10.21638/spbu23.2024.103

Статья поступила в редакцию 3 апреля 2024 г.;
рекомендована к печати 2 июля 2024 г.

Контактная информация:

Сунами Артем Николаевич — канд. полит. наук, доц.;
https://orcid.org/0000-0002-5418-8120, a.sunami@spbu.ru

Can Artificial Intelligence Be Xenophobic? Conflict and Ethical Challenges of Neural Network Development*

A. N. Sunami

St. Petersburg State University,
7–9, Universitetskaya nab., St. Petersburg, 199034, Russian Federation

For citation: Sunami A. N. Can Artificial Intelligence Be Xenophobic? Conflict and Ethical Challenges of Neural Network Development. *Vestnik of Saint Petersburg University. Philosophy and Conflict Studies*, 2024, vol. 40, issue 3, pp. 459–472. https://doi.org/10.21638/spbu17.2024.308 (In Russian)

The article examines the conflict and ethical challenges associated with the "the trouble with bias" of neural networks. Based on neural networks bias conceptions by Kate Crawford, Ezekiel Dixon-Roman and Luciana Parisi, a complex of natural language research for programming, the theory of cultural violence by Johan Galtung, the works by John Rawls and Will Kymlicka on the problem of justice and inequality, the author suggests that the experience of analysis of digital algorithms based on neural network technology has allowed us to form a corpus of ethical colored problems, the central of which is the "the trouble with bias". "The trouble with bias" is directly related to the phenomenon of "natural language", that is, a reflection of linguistic practices of the past and present, an average value of the digitized experience of mankind, which is the main resource for learning neural network algorithms. In this context, the article focuses to the stereotypes, power hierarchies, inequalities and discrimination encoded in natural language, which can be described in terms of "classification politics" and cultural violence. The article shows that that the ethical and conflict consequences of "the trouble with bias" can be reduced only if a particular social group receives the status of discriminated. On the basis of the identified substantial characteristics of the current neural network algorithms, the author concluded that there is a high danger of indoctrination of artificial intelligence algorithms with elements of xenophobia, exclusivity and "call-out culture".

Keywords: digitalization, neural network, the trouble with bias, natural language, inequality, classification politics, cultural violence.

References

1. Sunami, A. N. (2023), Ethics of "Digital Society": New Conflict or New Balance, *Vestnik of Saint Petersburg University. Philosophy and Conflict Studies*, vol. 39, iss. 3, pp. 544–556. https://doi.org/10.21638/spbu17.2023.311 (In Russian)

2. McCulloch, W. S. and Pitts, W. (1943), A logical calculus of the ideas immanent in nervous activity, *The Bulletin of Mathematical Biophysics. Kluwer Academic Publishers,* vol. 5, no. 4. pp. 115–133. https://doi.org/10.1007/BF02478259

3. Wiener, N. (1948), *Cybernetics or Control and Communication in the Animal and the Machine*, New York: Technology Press.

4. Hebb, D. O. (1949), *The Organization of Behavior. A Neuropsychological Theory*, New York: John Wiley & Sons.

5. Moravec, H. (1990), *Mind Children. The Future of Robot and Human Intelligence*, Harvard: Harvard University Press.

6. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman,D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A. W., Kavukcuo-

* The research was funded by the Russian Science Foundation grant no. 23-28-00220, https://rscf.ru/project/23-28-00220/.

glu, K., Kohli, P. and Hassabis, D. (2021), Highly accurate protein structure prediction with AlphaFold, *Nature*, vol. 596, pp. 583–589. https://doi.org/10.1038/s41586-021-03819-2

7. Lisovsky, A. L. (2020), Application of neural network technologies for management development of systems, *Strategicheskie resheniia i risk-menedzhment*, vol. 11, no. 4, pp. 378–389. https://doi.org/10.17747/2618-947X-2020-4-378-389 (In Russian)

8. Sozykin, A. V. (2017), An Overview of Methods for Deep Learning in Neural Networks, *Vestnik Iu-UrGU. Seriia: Vychislitel'naia matematika i informatika*, vol. 6, no. 3, pp. 28–59. https://doi.org/10.14529/cmse170303 (In Russian)

9. Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, A. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I. and Amodei, D. (2020), *Language Models are Few-Shot Learners*. https://doi.org/10.48550/arXiv.2005.14165

10. Du, M., He, F., Zou, N., Tao, D. and Hu, X. (2023), Shortcut Learning of Large Language Models in Natural Language Understanding, *Communications of the ACM*, vol. 67, no. 1, pp. 110–120. https://doi.org/10.1145/3596490

11. Yakovleva, L. I. (2007), The notion of common sense and the tradition which constructs it, *Vestnik Moskovskogo universiteta. Seriia 7: Filosofiia*, iss. 4, pp. 29–49. (In Russian)

12. Foucault, M. (1996), *The will to knowledge: Beyond knowledge, power and sexuality. Works of different years*, Moscow: Kastal' Publ. (In Russian)

13. Kejriwal, M. and Nagaraj, A. (2024), Quantifying Gender Disparity in Pre-Modern English Literature using Natural Language Processing, *Journal of Data Science*, vol. 22, no. 1, pp. 77–96. https://doi.org/10.6339/23-JDS1100

14. Gould, S. J. (1996), *The Mismeasure of Man*, New York: W. W. Norton & Co.

15. Crawford, K. (2023), *Atlas Of Ai Power, Politics And The Planetary Costs of Artificial Intelligence*, Moscow: AST Publ. (In Russian)

16. Galtung, J. (1990), Cultural Violence, *Journal of Peace Research*, vol. 27, no. 3, pp. 291–305.

17. Bowker, G. C. and Star, S. L. (1999), *Sorting Things Out: Classification and Its Consequences,* Cambridge, Mass.: MIT Press.

18. Dixon-Roman, E. and Parisi, L. (2020), Data capitalism and the counter futures of ethics in artificial intelligence. *Communication and the Public*, vol. 5, no. 3–4, pp. 116–121. https://doi.org/10.1177/2057047320972029

19. Wynter, S. (2001), Towards the sociogenic principle: Fanon, identity, the puzzle of conscious experience, and what it is like to be Black, in: Gomez-Moriana, A. and Duran-Cogan, M. (eds), *National identities and sociopolitical changes in Latin America*, Milton Pak: Routledge, pp. 30–66.

20. Rawls, J. (1995), *A theory of justice*, Novosibirsk: Izdatel'stvo Novosibirskogo universiteta Publ. (In Russian)

21. Kymlicka, W. (2010), *Contemporary Political Philosophy: An Introduction*, Moscow: Vysshaia shkola ekonomiki Publ. (In Russian)

22. Sunami, A. N. and Pavlova, E. V. (2024), The moral foundations of the politics under conditions of digital society flexibility: If human values no longer matter (part 1), *Political Expertise: POLITEX*, vol. 20, no. 1, pp. 33–44. https://doi.org/10.21638/spbu23.2024.103 (In Russian)

Received: April 3, 2024
Accepted: July 2, 2024

Author's information:

Artem N. Sunami — PhD in Political Sciences, Associate Professor;
https://orcid.org/0000-0002-5418-8120, a.sunami@spbu.ru